Journal of the Royal Statistical Society
SERIES C Applied Statistics

Appl. Statist. (2014)
63, *Part* 2, *pp.* 289–302

Linear mixed modelling for data from a double mixed factorial design with covariates: a case-study on semantic categorization response times

Jorge González B.,

Pontificia Universidad Católica de Chile, Santiago, Chile

Paul De Boeck,

Ohio State University, Columbus, USA, and KU Leuven, Belgium

and Francis Tuerlinckx

KU Leuven, Belgium

[Received July 2012. Final revision April 2013]

Summary. Linear mixed modelling is a useful approach for double mixed factorial designs with covariates. It is explained how these designs are appropriate for the study of human behaviour as a function of characteristics of individuals and situations and stimuli in the situations. The behaviour of subjects nested in types of individual responding to stimuli nested in types of stimuli defines a mixed factorial design. The inclusion of additional covariates of the observational units can help to explain the behaviour under study further. A linear mixed modelling approach for such designs allows a combined focus on fixed effects (general effects) and individual and stimulus differences in these effects. This combination has the potential to advance the integration of two different subdisciplines of psychology, general psychology and differential psychology, so that they can borrow strength from each other. An application is presented with semantic categorization response time data from a factorial design with age groups by word types and with age of acquisition as an additional covariate of the words. The results throw light on the processes underlying the effect of age of acquisition and on individual differences and word differences.

Keywords: Analysis of variance; Double mixed design; Mixed models; Random effects; Reaction time data

1. Introduction

Human behaviour is the topic of study in psychology and other social sciences. Behaviour is always the behaviour of a person in a context. This has been summarized symbolically in a famous expression by Lewin (1943): $B = f(P, E)$, where B, P and E stand for behaviour, person and environment or context respectively. Often the total context is too broad and more specific elements or aspects from the context are focused on. A generic term for the rather specific elements or aspects from the context is stimulus. Examples of stimuli are other individuals (social behaviour), commercial products (consumer behaviour) and words (linguistic behaviour). The study of behaviour consists of relating behaviour to characteristics of the individuals and of the context and stimuli therein. Roughly speaking, the characteristics can be categorical or

Address for correspondence: Paul De Boeck, Department of Psychology, Ohio State University, Lazenby Hall, 1827 Neil Avenue, Columbus, OH 43210, USA.
E-mail: deboeck.2@osu.edu

© 2013 Royal Statistical Society 0035–9254/14/63289

continuous. In the application, the response time of semantic categorizations of nouns will be studied as a function of age group (young *versus* old adults) and type of object denoted by nouns (natural object *versus* artefact object). The nouns are the stimuli. In other applications the stimuli can be individuals, cognitive problems, emotional cues, etc. Given that categorical characteristics of individuals and stimuli are focused on, it is evident that multiple instantiations of each category should be included.

The principles of random sampling and random assignment, as a basis for generalization and for causal inference respectively, apply in a double sense when a double mixed factorial design is used. Random sampling for the double mixed factorial design refers to both sides of the design. For the application, these are individuals and nouns. Random assignment means that the elements are randomly assigned to the levels of the factors. For the application, this would mean that the individuals are randomly assigned to age groups and that the nouns are randomly assigned to the natural object category and the artefact category. This is of course not possible. Age and word type are not manipulable. Hence, the application is an observational but not an experimental study. Although under certain conditions causal inference is also possible from observational studies (McGue *et al.*, 2010), we do not plan to draw causal conclusions. When random assignment is not possible, the random-sampling principle refers to sampling from the levels of the factors, in our case from the two age groups (two subpopulations of people) and from the two word types (two subpopulations of nouns). As will be described in Section 2.2, random sampling was not fully realized in the application.

Independently of the observational or experimental nature of the study, the designs that are described are double mixed factorial designs. The categories and the experimental conditions define the levels of fixed factors and the sampled units; the subjects and the stimuli define the random factors. Because the mixed nature of the design (fixed and random) applies to the person side and to the stimulus side of the design, we use the terms *double* and *mixed*. The term *factorial* is added because it is commonly used in the social sciences and in an analysis-of-variance context for categorical explanatory variables. This format of the study design is nicely in line with the widely accepted understanding that behaviour is a function of the person and the environment as stated in the above-mentioned expression $B = f(P, E)$ (Lewin, 1943).

Not only nominal category variables are of interest in the study of human behaviour but also continuous aspects of individuals and stimuli. In the application, the age at which a word is acquired (on average) will be investigated as a potential factor in a semantic categorization task. In line with the terminology for analysis of covariance, the term covariate will be used for such continuous explanatory variables. When combined with the study design as described thus far, the resulting design will be called a double mixed factorial design with covariates. From the modelling perspective that we shall describe next, the (categorical) factors and the covariates are all covariates, or factors, or whatever term one would prefer. Independent of their categorical or continuous nature and whether balanced or not, they are the explanatory variables (independent variables) for a given response variable (dependent variable).

Whether the study design is observational or experimental, the statistical models that can be applied are the same. If the response variable is continuous, linear mixed modelling would be appropriate, with fixed effects of the categorical factors (e.g. age group and word type) and the covariates (e.g. age of acquisition) and with random effects due to the random factors (subjects and stimuli) nested within the levels of the fixed factors. This is an instance of linear mixed modelling for cross-classified data. This conceptually rather simple study design leads to a large number of possible model specifications, especially regarding the random effects. For example, for the random-person effects, the structure can be as simple as a random intercept with equal variance in the two age groups, but it can also be as complex as including a random slope and

a random intercept with a different covariance structure depending on the age group. The full complexity of random-effect structures is described in Section 3.

The random-effect structure is of importance for mainly two reasons. First, the statistical test results for the fixed effects may depend on the assumed structure of the random effects. Practices largely differ in this regard, as we may derive from an overview by Barr *et al.* (2013), who recommended a maximal structure, but the design that they considered is simpler than in the present application. Second, although the random effects are often considered as a background that one needs to take into account for a proper test of the fixed effects, we believe that they are often also a valuable source of information. In psychology, a division can be noticed between the study of general phenomena, as in cognitive and social psychology, where one is interested in the first place in fixed effects, and the study of individual differences, as in the domain of intelligence and personality, where one is interested in latent traits. See Cronbach (1957) for a discussion. Mixed modelling offers an opportunity to integrate these two approaches to some extent, by focusing on general (fixed) effects and individual differences (random effects) at the same time. With a mixed model for cross-classified data one can even go one step further and also investigate stimulus differences and not just individual differences in how they relate to people's behaviour.

The paper is organized as follows. We first give a description of the research questions and the data set in Section 2. In Section 3 we describe the modelling approach and how aspects of the models are linked with the research questions. In Section 4 the data are analysed on the basis of the modelling approach proposed. The paper concludes with a discussion and final comments in Section 5.

The data that are analysed in the paper and the R programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Research questions and data

2.1. Research questions

The data are from a study by De Deyne and Storms (2007) focusing on how the age of acquisition of words is related to the response time in a semantic categorization task. The logarithm of the response time is the response variable. The design is a 2×2 design: age group A (young or old) by word type T (artefact or natural), with subjects s nested in the two levels of A, and with words w nested in the two levels of T. All subjects are crossed with all words. Also three continuous covariates are available for the words: word frequency freq, word familiarity fam and age of acquisition aoa. On the basis of a preliminary analysis to be reported in Section 4, only the age of acquisition is used for a further analysis of the data. No subject covariates are available in this data set and we have no information on the precise age of each subject. After the preliminary step in the analysis, it turns out that age group, word type and the age of acquisition are the only explanatory variables that will be used for further analysis.

Age of acquisition is considered an important characteristic of words in psycholinguistics. It is related to various types of linguistic behaviour (e.g. Bonin *et al.* (2004)). We shall look at the data from two possible theoretical perspectives related to age of acquisition. The first perspective is that the age of acquisition determines the cumulative frequency of encounters with the word in question and that a higher cumulative encounter frequency leads to faster response times in semantic tasks. On the basis of this cumulative frequency hypothesis we expect that the effect of differences between words in age of acquisition is smaller for older subjects than for younger subjects. The expectation is based on the assumption that the effect of an increasing number

of encounters diminishes with the number of encounters and thus with age (as in the law of diminishing marginal returns). The second perspective is that age of acquisition is associated with the structure of the lexicon at the time of acquisition, and that a younger age of acquisition is associated with a more central position in the lexicon and therefore leads to faster response times in semantic tasks. On the basis of this structure-of-lexicon hypothesis, a stronger effect of age of acquisition may be expected for natural objects (e.g. animals and plants) than for artefact objects (e.g. containers, drugs, vehicles, buildings and gadgets), because centrality in the lexicon is less relevant when the semantic structure is loose and variable, such as for artefacts. The semantic structure of artefacts is more scattered, more complex and less stable than the semantic structure of natural objects. Just to give one example, technology provides us with ever new artefacts and new categories of artefacts, e.g. sensors of all kinds. From a structure-of-lexicon point of view, the advantage of an earlier acquisition is therefore less clear for artefact words.

In sum, on the basis of the cumulative frequency hypothesis, we may expect an interaction of age of acquisition with age group. On the basis of the structure-of-lexicon hypothesis, we may expect an interaction of age of acquisition with word type. When both principles apply, the corresponding two interactions would show. When both principles play a role, it is also possible that one moderates the other, resulting in a second-order interaction of age of acquisition with age group and word type.

The authors of the original study (De Deyne and Storms, 2007) were primarily interested in the fixed effects of age of acquisition, and they used linear regression per age group and per subject. We are also interested in the fixed effects of the design factors and in individual differences and word differences, and thus in random-subject and random-word effects, because they are relevant for the kind of underlying processes. When the individual differences with respect to natural objects are of a different kind from individual differences with respect to artefacts, then this is an indication for differences in the cognitive processes for the two word types, and it would be in line with our conjecture about the difference between the two lexicons. When the word differences are of a different kind depending on the age of the subjects, then this is an indication for age differences in how the words are processed.

The specific research questions are grouped into two categories. The first category concerns the general effects (fixed effects). The second category concerns the varying effects (random effects).

First, from the two perspectives, the cumulative frequency perspective and the structure-of-lexicon perspective, we are interested in the interaction of age of acquisition with age group and word type, and therefore also in the lower order effects.

(a) Do young and old people differ in the speed of the semantic categorization task?
(b) Does the speed of the semantic categorization task depend on the type of word?
(c) Are word type differences the same in both age groups?
(d) What is the effect of age of acquisition and does it depend on age group, word type and their combination?

Second, because we want to combine the two approaches, for general phenomena (fixed effects) and for individual differences and word differences (random effects), and because these differences are relevant for the nature of the underlying processes, we have also research questions regarding the random effects.

(e) Are possible individual differences in response time general or do they depend on the word type, and, when they depend on the word type, how are the individual differences correlated?

(f) Are possible word differences in response time general or do they depend on the age group, and, when they depend on the age group, how are the word differences correlated?

2.2. Data

2.2.1. Subjects

The participants were 21 young adults from age 18 to 20 years, and also 21 older adults from age 52 to 56 years. They were the only participants in the study. The young adults were students who participated for course credits. The older adults were volunteers knowing that they would receive a movie theatre film ticket for their participation. They all had a higher education degree so their educational level was roughly equivalent to that of the university students. All participants were native Dutch speakers. It is quite common in studies in psychology that students participate for course credit and that other participants are volunteers who receive a small reward. Because the sampling mechanism is not random, it is not clear whether possible age differences can be generalized.

2.2.2. Words

The subjects were presented with 160 nouns that had to be categorized as referring to a natural object or an artefact object and the response times were registered. The nouns were selected from available norm lists obtained from studies described in De Deyne and Storms (2007). The norm lists concern three aspects of the nouns: familiarity, age of acquisition and frequency. Subjects from these earlier studies have rated words on their *familiarity* on a five-points scale: 'The word has been encountered or used 1—never, 2—almost never, 3—sometimes, 4—often, 5—very often'. The subjects from these earlier studies were also requested to estimate the *age of acquisition*, i.e. the age at which they estimated that they first had learned the words. The procedure was fully described and also validated by Ruts *et al.* (2004). *Frequency* norms for these words are available from the international CELEX database (Baayen *et al.*, 2003). The 160 words were selected such that the number of letters was in the range from 3 to 10, such that it was clear to which of the two categories a word belongs ('artefact' or 'natural'), and such that there is an equal number of both word categories. From the word selection as described, it is not clear whether possible differences between the two word categories can be generalized.

2.2.3. Actual data for the analysis

The participants in the study were presented with all 160 words and had to classify each word in one of two categories: natural or artefact. The order of presentation was randomized within subjects. Response times in milliseconds were registered. The data to be analysed are the natural-log-transformed response times of correct responses. This makes the data in fact slightly unbalanced. De Deyne and Storms (2007) reported that, after removing 14 words with an error rate larger than 40%, the error rates drop to 3.7% for the younger group and 3.5% for the older group. Among the remaining words there are 70 denoting natural objects and 76 denoting artefact objects.

3. Modelling

3.1. Linear mixed models

As noted by Baayen *et al.* (2008), Freeman *et al.* (2010), Jaeger (2008) and Quené and Van den Bergh (2008) in the field of psycholinguistics, models known in different contexts as hierarchical

linear models, multilevel models or linear mixed models (e.g. Raudenbush and Bryk (2001), Verbeke and Molenberghs (2000), Pinheiro and Bates (2000), Snijders and Bosker (2011) and Gelman and Hill (2007)) are highly useful for an analysis of responses to stimuli. In linear mixed modelling the response variable is modelled as a function of fixed effects and random effects of a set of explanatory variables and their interactions. The response data are denoted as Y_{sw}, with s as an index for subjects and w as an index for words. The equation that we start from to explain the modelling approach is

$$Y_{sw} = \mathbf{X}'_{sw}\boldsymbol{\beta} + \mathbf{W}'_{w}\mathbf{u}_{WA[s]} + \mathbf{S}'_{s}\mathbf{u}_{ST[w]} + \varepsilon_{sw} \tag{1}$$

where $\varepsilon_{sw} \sim N(0, \sigma^2_\varepsilon)$ is the residual. The other three terms on the right-hand side of equation (1) are explained in what follows. For the two types of random effects, the most general formulation will be described, but this general formulation will later be constrained in different ways.

The first term corresponds to the *fixed part* of the model. $\mathbf{X}_{sw}$ is the vector of explanatory variables with $\boldsymbol{\beta}$ as fixed effects. It entails the constant predictor, the fixed design factors age group (denoted by A) and word type (denoted by T), the three covariates and the products between all these (for the interactions). As will be explained, in our analysis word frequency and word familiarity will be omitted after a preliminary analysis. Age group and word type are dummy coded as follows: young, 1; old, 0; artefact, 1; natural, 0. For example, for a subject from the young group responding to a noun for a natural object, given that the age of acquisition was 7 years, the entries of the vector would be 1 for the intercept, 1 for A (young), 0 for T (natural), 7 for aoa, 0 for $A \times T$, 7 for $A \times$ aoa, 0 for $T \times$ aoa, and 0 for $A \times T \times$ aoa: $(1, 1, 0, 7, 0, 7, 0, 0)'$.

The second term corresponds to the *random-subject part* of the model. $\mathbf{W}_w$ is the vector of explanatory word variables with random-subject effects $\mathbf{u}_{WA[s]}$. Subscript W indicates that these are effects of $\mathbf{W}_w$, whereas subscript $A[s]$ indicates that they are random-subject effects and that the subjects are nested within the levels of A. The distribution of the random effects is $\mathbf{u}_{WA[s]} \sim N(0, \boldsymbol{\Sigma}_{WA})$, the structure of individual differences in the study. The random-subject effects can have a different structure depending on the level of A. For the vector $\mathbf{W}_w$ we shall use a different coding from that for $\mathbf{X}_{sw}$. Instead of using a random intercept and a random slope for the difference between the two word types, an alternative but equivalent coding is used, with a first random-subject effect for artefact words and a second for natural. This is because we are interested in the individual differences per word type. Therefore, the vector $\mathbf{W}_w$ has a binary entry for artefact and another for natural. For the same example as before, the coding would be 0 for $T \equiv$ artefact, 1 for $T \equiv$ natural: $(0, 1)'$. In the analysis we shall also briefly consider the possibility of random-subject effects of aoa and how they may depend on A and T. We chose not to complicate the general model formulation with the inclusion of these effects. They will be briefly explained later.

The third term corresponds to the *random-word part* of the model. $\mathbf{S}_s$ is the vector of explanatory subject variables with random word effects $\mathbf{u}_{ST[w]}$. Subscript S indicates that these are effects of $\mathbf{S}_s$, whereas subscript $T[w]$ indicates that they are random-word effects and that the words are nested within the levels of T. The distribution of the random effects is $\mathbf{u}_{ST[w]} \sim N(0, \boldsymbol{\Sigma}_{ST})$, the structure of word differences in the study. The random-word effects can have a different structure depending on the level of T. For the vector $\mathbf{S}_s$ we shall again use a different coding from that for $\mathbf{X}_{sw}$, with a first random-word effect for young and a second for old. This is because we are interested in word differences per age group. The vector $\mathbf{S}_s$ for the same example as before would be 1 for $A \equiv$ young and 0 for $A \equiv$ old: $(1, 0)'$.

The model as described is on the one hand a standard linear mixed model. On the other hand, the type of modelling and the effects that we are interested in are not commonly considered in psychology and the social sciences although one might expect that they are on the basis of the $B =$

$f(P, E)$ idea. From a statistical point of view, a between-subject factor (age group) and a within-subject factor (word type) are formally the same in a crossed design, but they are quite different for the psychological interpretation of the effects. Although within-subject factors are commonly used, it is rare to have random effects nested in the levels of such a factor, whereas within-category variation of stimuli is in line indeed with behavioural reality. Although it is evident that such variation exists, the data are often aggregated across stimuli so that the variation is invisible and left unexplained. The use of a covariate as in the application can contribute to such an explanation. Furthermore, commonly random-subject effects function as background variation in a model for data to investigate general phenomena that are tested with fixed effects, whereas we are interested in more. Focusing on random-subject effects (individual differences) and their structure in a study as presented here is an opportunity to combine two different subdisciplines, more specifically general psychology and the psychology of individual differences, and for one subdiscipline to borrow substantive strength from the other subdiscipline. In addition, we are also interested in the random-word effects, as explained. In the next two subsections, the more specific fixed effect and random-effect parts of the model will be linked with the general model of equation (1) on the one hand and with the research questions on the other hand.

3.2. Fixed effects

The research questions (a)–(d) can be answered on the basis of the estimated $\boldsymbol{\beta} = (\beta_0, \beta_A, \beta_T, \beta_{AT}, \beta_a, \beta_{aA}, \beta_{aT}, \beta_{aAT})'$, referring to the intercept, to the effect of the dummy variable for A (β_A, question (a)), T (β_T, question (b)), $A \times T$ (β_{AT}, question (c)), aoa (β_a, question (d)) and its products with A and T ($\beta_{aA}, \beta_{aT}, \beta_{aAT}$, question (d)). The remaining questions concern the random effects.

3.3. Structure of random-subject effects

Four different kinds of structures will be considered. The upper part of Table 1 gives a summary. The four specifications for the random-subject effects concern the $\mathbf{u}_{WA[s]}$:

(a) one general random effect θ_s, the random intercept, with $\theta_s \sim N(0, \tau_\theta)$ (we shall call this first structure *unidimensional homoscedastic*; it implies that there is only one source of individual differences; in terms of the general model of equation (1), the vector $\mathbf{W}_w$ can be reduced to $\mathbf{W}_w \equiv (1)$, and the random-effect vector to $\mathbf{u}_{WA[s]} = (\theta_s)$);

(b) a different variance depending on the age group, $\theta_{A[s]} \sim N(0, \tau_{\theta_A})$ (we shall call this second structure *unidimensional heteroscedastic*; it implies that there is only one source of individual differences within each age group, but that the variance can differ between the age groups; in terms of the general model of equation (1), $\mathbf{W}_w \equiv (1)$, and $\mathbf{u}_{WA[s]} = (\theta_{A[s]})$);

Table 1. The four assumptions for random-subject effects and random-word effects

	Homoscedastic	*Heteroscedastic*
Subjects		
Unidimensional	$\theta_s \sim N(0, \tau_\theta)$	$\theta_{A[s]} \sim N(0, \tau_{\theta_A})$
Multi-dimensional	$\boldsymbol{\theta}_s \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$	$\boldsymbol{\theta}_{A[s]} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\theta_A})$
Words		
Unidimensional	$\delta_w \sim N(0, \tau_\delta)$	$\delta_{T[w]} \sim N(0, \tau_{\delta_T})$
Multi-dimensional	$\boldsymbol{\delta}_w \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\delta)$	$\boldsymbol{\delta}_{T[w]} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\delta_T})$

(c) a different random effect depending on the word type, $\boldsymbol{\theta}_s \sim N(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$ (we shall call this third structure *multi-dimensional homoscedastic*; it implies that the individual differences differ for natural and artefact, but that their covariance structure does not depend on the age group; in terms of the general model of equation (1), $\mathbf{W}_w = (1,0)'$ for artefact and $\mathbf{W}_w = (0,1)'$ for natural, and $\mathbf{u}_{WA[s]} = (\theta_{1s}, \theta_{2s})'$);

(d) a different random effect depending on the word type and a different covariance matrix depending on the age group, $\boldsymbol{\theta}_{A[s]} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\theta_A})$ (we shall call this second structure *multi-dimensional heteroscedastic*; it implies that the individual differences differ for natural kinds and artefacts and that the covariance matrix of the corresponding random effects depends on the age group; in terms of the general model of equation (1), $\mathbf{W}_w = (1,0)'$ for artefact and $\mathbf{W}_w = (0,1)'$ for natural, and $\mathbf{u}_{WA[s]} = (\theta_{1A[s]}, \theta_{2A[s]})'$).

Research question (e) can be answered by comparing the goodness of fit of the four covariance structures just described and presented in the upper part of Table 1.

3.4. Structure of random-word effects

In an analogous way four structures can be defined for the random-word effects (Table 1):

(a) unidimensional homoscedastic,

$$\mathbf{S}_s = (1), \qquad \mathbf{u}_{ST[w]} = (\delta_w), \qquad \delta_w \sim N(0, \tau_\delta);$$

(b) unidimensional heteroscedastic,

$$\mathbf{S}_s = (1), \qquad \mathbf{u}_{ST[w]} = (\delta_{T[w]}), \qquad \delta_{T[w]} \sim N(0, \tau_{\delta_T});$$

(c) multi-dimensional homoscedastic,

$$\mathbf{S}_s = (1,0)', \text{for young}, \qquad \mathbf{S}_s = (0,1)', \text{for old}, \qquad \mathbf{u}_{ST[w]} = (\delta_{1w}, \delta_{2w})', \qquad \boldsymbol{\delta}_w \sim N(0, \boldsymbol{\Sigma}_\delta);$$

(d) multi-dimensional heteroscedastic,

$$\mathbf{S}_s = (1,0)', \text{for young}, \quad \mathbf{S}_s = (0,1)', \text{for old}, \quad \mathbf{u}_{ST[w]} = (\delta_{1T[w]}, \delta_{2T[w]})', \quad \boldsymbol{\delta}_{T[w]} \sim N(0, \boldsymbol{\Sigma}_{\delta_T}).$$

Research question (f) can be answered by comparing the goodness of fit of the four covariance structures just described and presented in the lower part of Table 1. Combined with the four possible random-subject structures, there are in total 16 ($=4 \times 4$) possible structures.

4. Application

4.1. Plan for analysis

The analysis will be conducted through different steps. In a preliminary step it will be checked whether also other available covariates than aoa (freq, fam) should be included in the model. Next, it will be investigated which of the 4×4 random-effect structures is the best fitting one while including all fixed effects of interest. As explained, the structure of the random effects is itself of substantive interest and not only as an assumed structure for testing the fixed effects. The modelling will then be continued with the best structure for the random effects, but, we shall also check whether the results would change depending on the choice that is made. Because the interactions of age group and word type with age of acquisition turn out to be significant, also the random effects of this covariate are included in the model, following a suggestion by Baayen and Milin (2010).

4.2. Preliminary analysis

Because freq, fam and aoa are correlated, we want to investigate whether they all three should be included in the model. In this preliminary step, the data are analysed with two models, both with all three covariates and for each of these covariates also their interaction with age group and word type. The two models each correspond to an extreme set of assumptions: one model with unidimensionality and homoscedasticity for subjects and words (minimal structure), and another with multi-dimensionality and heteroscedasticity for subjects and words (maximal structure) (Barr *et al.*, 2013). Both models are estimated with maximum likelihood by using the `lmer` function from the `lme4` package in R (Bates *et al.*, 2011). For both models only the effects with age of acquisition involved turned out to be significant. Furthermore, omitting the effects that are related to the other two covariates did not lead to a significant likelihood ratio test, whereas omitting the effects related to age of acquisition did lead to a significant decrease in goodness of fit based on the likelihood ratio test.

Following these results, we decided to continue with all main effects and interaction effects of age group, word type and age of acquisition, aoa. For the estimation, maximum likelihood and restricted maximum likelihood were used. The conclusions are very similar for both estimation methods, and therefore only the maximum likelihood results are reported.

4.3. Random-effects structure

Table 2 shows the number of parameters, the log-likelihood, the Akaike information criterion (AIC) (Akaike, 1974) and the Bayesian information criterion (BIC) (Schwarz, 1978) of the models that correspond to the 4×4 random-effect structures. Each of the models includes also the fixed effects of A, T and aoa, and their interactions. Following the AIC and the BIC, the multi-dimensional homoscedastic structure is the best fitting one for the random-subject-

Table 2. Goodness of fit of the 16 models†

Model	*Number of parameters*	*AIC*	*BIC*	*Log-likelihood*
M-s1o-w1o	11	−4455	−4383	2239
M-s1o-w1e	12	−4454	−4375	2239
M-s1o-w2o	13	−4467	−4382	2246
M-s1o-w2e	16	−4475	−4370	2253
M-s1e-w1o	12	−4453	−4374	2239
M-s1e-w1e	13	−4451	−4366	2239
M-s1e-w2o	14	−4465	−4373	2246
M-s1e-w2e	17	−4473	−4361	2253
M-s2o-w1o	13	−4498	*−4412*	2262
M-s2o-w1e	14	−4496	−4404	2262
M-s2o-w2o	15	−4510	*−4412*	2270
M-s2o-w2e	18	*−4519*	−4400	2277
M-s2e-w1o	16	−4496	−4391	2264
M-s2e-w1e	17	−4494	−4383	2264
M-s2e-w2o	18	−4509	−4391	2272
M-s2e-w2e	21	−4517	−4379	2279

†The best fitting models are indicated in italics. The model labels have two components: one for random-subject effects (s) and one for random-word effects (w). The component symbol (s or w) is followed by the number of dimensions (1 or 2) and an indication of homoscedastic (o) or heteroscedastic (e).

Table 3. Estimates of the fixed effects, standard errors and *p*-level, for age group *A*, word type *T* and age of acquisition aoa, and estimates of the random-effect distributions (standard deviation of the final model)

Effect	*Estimate*	*Standard error*	*p-level*	*Standard deviation*
Intercept	6.8853	0.0260	<0.001	
Age group—*A* (young)	−0.1678	0.0340	<0.001	
Word type—*T* (artefact)	0.1854	0.0256	<0.001	
Age of acquisition—aoa	0.0094	0.0016	<0.001	
$A \times T$	0.0264	0.0323	>0.10	
$A \times$ aoa	0.0100	0.0033	<0.01	
$T \times$ aoa	−0.0086	0.0018	<0.001	
$A \times T \times$ aoa	−0.0108	0.0038	<0.01	
Random-subject artefact				0.0875
Random-subject natural†				0.0932
Random-subject aoa old				0.0014
Random-subject aoa young				0.0074
Random-word				0.0621
Residual				0.1490

†Correlation with random-subject artefact 0.885.

effect structure, independent of the structure for the words. The result is less clear for the word structure. The AIC indicates the multi-dimensional heteroscedastic structure, whereas there are two models with an equally lowest BIC value: the unidimensional and multi-dimensional homoscedastic structures. A possible explanation for the lack of clarity is that the data are less informative regarding the word structure because the word differences are based on only two times 21 subjects whereas the individual differences are based on 70 and 76 words. The results provide evidence for the individual differences being of a somewhat different kind depending on the word type, although the correlation is as high as 0.880. In contrast, there is at best only mixed evidence in favour of a two instead of just one random-word effect. Larger samples of both age groups may be required for a stronger inference. We continue with the most simple of the two best BIC models, model M-s2o-w1o from Table 2, but it will also be checked whether this choice makes a difference.

It can be concluded from the estimated random-effect structures that the individual differences for natural object nouns and artefact nouns are highly correlated but also different. It seems therefore that the underlying processes for the semantic categorization task are of a somewhat different kind, in line with the assumed difference in structure between the two lexicons.

In the selected model M-s2o-w1o and in all other 15 models from Table 2, all interaction effects with aoa involved were significant. The estimates were all highly similar as may be expected (Zeger *et al.*, 1988), but also the *p*-values are not really affected in this case. We do not report the exact estimates since they are highly similar to those shown in Table 3.

Following a suggestion by Baayen and Milin (2010) to include also random effects of an interacting covariate, we explored whether the effect of aoa was varying across subjects. Because of the interactions of aoa with *A* and *T*, and with $A \times T$, the most general estimated model was one with a random-aoa slope in each of the four cells of the $A \times T$ design. Also all simpler structures for the random-aoa slopes were estimated, including one with an overall random slope of aoa, each time with and without correlations with the other random-subject effects. It would lead us too far to spell out all these different random-effect models. For the most general

of the correlated models, no convergence was reached. On the basis of the AIC and BIC values of all these variants, the best model was one with a random-aoa slope only in the group of young adults and independent of the other random-subject effects. The AIC and BIC values of the model were −4510 and −4418 respectively. The improvement compared with model M-s2o-w1o is rather small (see Table 2). Because it is rather uncommon to use random effects limited to one group, and because we did not have a hypothesis regarding the absence of slope variation in the older group, we decided to continue with the model that does have a random slope in both age groups. This model has an AIC and BIC of −4510 and −4411 respectively for the younger and older age groups. Table 3 presents the estimates for the model in question and it can be noted that the standard deviation for the older group is much smaller than for the younger group and also extremely small in an absolute sense: 0.0014 (old) compared with 0.074 (young). All other estimates are highly similar to those of the model without a random slope. The residual standard deviation SD of the model without the random slope was reduced from 0.1498 to 0.1490.

4.4. Discussion of the final model

To interpret the fixed effects in Table 3, Fig. 1 is constructed. It shows the expected natural-log-transformed response times in the four cells of the design, for two reference points on the age of word acquisition continuum, and for all the random-effect values set equal to 0. The two ages of word acquisition, 5 years and 20 years, cover a very broad range (mean 9.78; SD 5.55). The expected values from Fig. 1 are also shown in Table 4 under the column heads 'Mean'. The effect of age of acquisition seems limited to natural objects and is stronger for younger subjects than for older subjects. It is also clear from Fig. 1 that in general the younger subjects respond

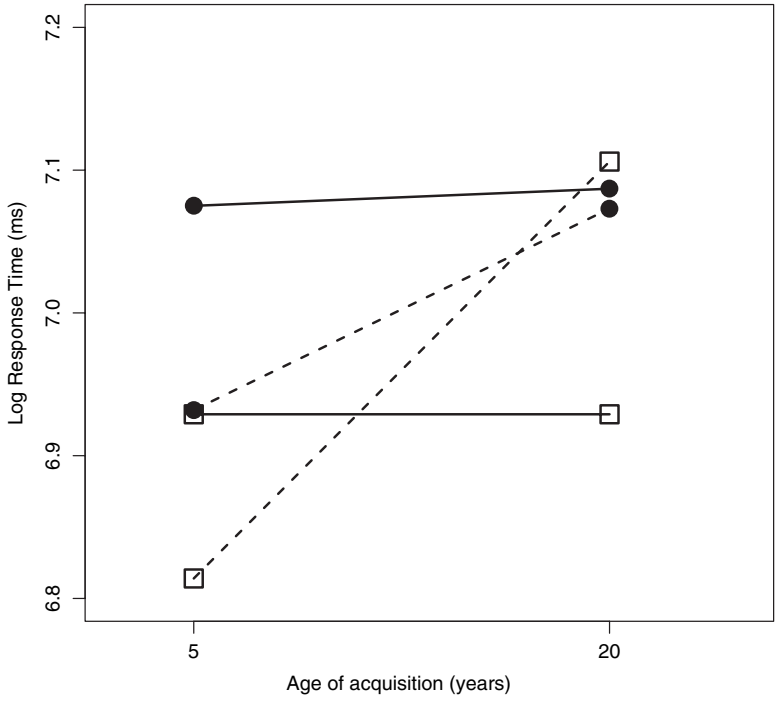


Fig. 1. Model-based natural logarithm response time in the four cells of the design as a function of age of acquisition: ———, artefact; – – –, natural; ●, old; □, young

Table 4. Model-based natural logarithm response times as a function of age of acquisition depending on age group and word type with an indication of the range (from −2 SD to 2 SD) of the random slopes of age of acquisition

	Response times (ms) for artefact objects			*Response times (ms) for natural objects*		
	−2 SD	*Mean*	*2 SD*	*−2 SD*	*Mean*	*2 SD*
Young						
aoa = 5 years old	6.855	6.929	7.003	6.740	6.814	6.888
aoa = 20 years old	6.633	6.929	7.225	6.810	7.106	7.402
Old						
aoa = 5 years old	7.061	7.075	7.089	6.918	6.932	6.946
aoa = 20 years old	7.031	7.087	7.143	7.017	7.073	7.129

faster than the older subjects. To have a view on the individual differences in the effect of age of acquisition, Table 4 shows the range from −2 SD to 2 SD around the fixed-effect-based means. For natural objects the effect is positive across the whole range of individual differences, but for artefacts the effect is reversed in the lower range of the individual differences (see −2 SD for artefact). Because the model with random-aoa slopes is not much better than the model without (M-s2o-w1o), we do not draw strong conclusions from this finding. That the possible individual differences in the effect are concentrated in the younger age group is not surprising since the distance between the subjects' age and the age of acquisition is much smaller than in the older age group.

It can be concluded from these results that both the cumulative frequency perspective and the structure-of-lexicon perspective are supported by the data. The interaction of age of acquisition with age group supports the cumulative frequency hypothesis, whereas the interaction of age of acquisition with word type supports the structure-of-lexicon hypothesis. The second-order interaction ($A \times T \times$ aoa) from Table 3 suggests that the two principles are at work and that they interact.

5. General discussion

With respect to the substantive research questions, one can infer from the application that age of acquisition of words is important for semantic categorization and that both the cumulative frequency and the structure of lexicon seem to play a role. Another interesting finding is that the individual differences with respect to the categorization of natural objects and artefact objects are of a slightly different kind, suggesting, in line with our theorizing, that the processes are somewhat different. We also found some evidence for individual differences in the effect of age of acquisition, mainly in the younger age group.

Although the design and data analysis of the application are very much in line with the general basic principle that behaviour depends on the person and the stimuli in the environment of the behaving person ($B = f(P, E)$), it is still rather uncommon to see this approach fully implemented. As illustrated with the application, the approach makes it possible

(a) to study individual differences in the context of general effects so that the individual differences derive meaning from the general theory and

(b) to supplement the study of general effects with individual and stimulus differences regarding these general effects.

This is a way to integrate two otherwise rather isolated subdisciplines.

The approach has become more popular in psycholinguistics, as illustrated by an overview in Barr *et al.* (2013) and among others by Baayen *et al.* (2008) and Baayen and Milin (2010). Also, in quite a different domain from the social sciences, a design similar to a double mixed factorial design was considered by Stanley (1961) in which a sample of raters (college students) provide ratings for a sample of ratees (public figures). Both random factors, raters and ratees, are nested in fixed factors. Stanley (1961) called such a design a 'double nested design'. In another example, from educational sciences, a sample of students is measured once in elementary school and another time in junior high school; the two crossed random factors are elementary schools and junior high schools (and the students are nested in the cells) (Shi *et al.*, 2010). The crossing is the result of the fact that the students at the end of elementary school spread out to different junior high schools. The type of elementary school and the type of junior high school (e.g. public *versus* private) are the fixed factors. Because students and schools are different random-factor levels, this is a multilevel double mixed design.

An important issue in the analysis of data from similar but also simpler designs (e.g. with only random-subject effects) is how one can deal with the random-effect structures. Barr *et al.* (2013) gave an overview of approaches, such as using exclusively random intercepts, and using forward and backward model selection for random-effect structures. Barr *et al.* (2013) concluded that there is a large variety of approaches without an accepted standard. One can find valuable considerations in the literature such as guarding against anticonservative conclusions by using a minimal random-effect structure based on theoretical considerations (Pinheiro and Bates, 2000), and such as fitting only one *a priori* chosen model (Harrell, 2001) as a way to guard against opportunistic conclusions. Barr *et al.* (2013) proposed a maximal random-effect structure and showed with simulation studies that it is a valuable approach. Given this variety of approaches we wanted to check the robustness of the fixed effect tests across all possible random-effect structures. This would perhaps be too complex an approach if the number of structures is much larger, and a problem remains when the test results do not converge. However, in such a case, one should perhaps not draw strong conclusions anyway.

It was not our aim to propose a general solution for the problem. It was instead our purpose to illustrate a possible approach: one that is inspired by an interest in the random effects also for reasons other than testing fixed effects. These other reasons are substantive and inspired by the possibility of borrowing strength for one subdiscipline (differential psychology) from another subdiscipline (general psychology) and vice versa.

Several aspects of the linear mixed modelling analysis are not discussed here because we focused on a case-study. First, the normal distribution for random effects may be considered a limitation. Alternatives are described in the literature, such as the use of skew normal distributions or Dirichlet-process-based distributions (Jara *et al.*, 2008; Arellano-Valle *et al.*, 2005; Mukhopadhyay and Gelfand, 1997). Second, it might be worth checking whether the effects of covariates are non-linear so that, especially in combination with a larger number of covariates and interactions, moving to a generalized additive model (e.g. Wood (2011)) may be worth considering. Third, a Bayesian approach of linear mixed modelling is an attractive alternative for the maximum likelihood and restricted maximum likelihood approaches.

Acknowledgements

We thank the Joint Editor and two referees for their very helpful comments and suggestions. We

especially thank the Guest Associate Editor for an extensive and extremely helpful review. We thank Simon De Deyne and Gert Storms for kindly providing us with the data.

References

Akaike, M. (1974) A new look at the statistical model identification. *IEEE Trans. Autom. Control*, **19**, 716–723.
Arellano-Valle, R., Bolfarine, H. and Lachos, V. (2005) Skew-normal linear mixed models. *J. Data Sci.*, **3**, 415–438.
Baayen, R., Davidson, D. and Bates, D. (2008) Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.*, **59**, 390–412.
Baayen, R. and Milin, P. (2010) Analyzing reaction times. *Int. J. Psychol. Res.*, **32**, 12–28.
Baayen, R., Piepenbrock, R. and van Rijn, H. (2003) The CELEX lexical database. *Compact Disc*. Linguistic Data Consortium. University of Pennsylvania, Philadelphia.
Barr, D., Levy, R., Scheepers, C. and Tily, H. (2013) Random effects structure for confirmatory hypothesis testing: keep it maximal. *J. Mem. Lang.*, **68**, 255–278.
Bates, D., Maechler, M. and Bolker, B. (2011) lme4: linear mixed-effects models using S4 classes. *R Package Version 0.999375-42*.
Bonin, P., Barry, C., Meot, A. and Chalard, M. (2004) The influence of age-of-acquisition in word reading and other tasks: a never ending story? *J. Mem. Lang.*, **50**, 456–476.
Cronbach, L. (1957) The two disciplines of scientific psychology. *Am. Psychol.*, **12**, 671–684.
De Deyne, S. and Storms, G. (2007) Age-of-acquisition differences in young and older adults affect latencies in lexical decision and semantic categorization. *Acta Psychol.*, **124**, 274–295.
Freeman, E., Heathcote, A., Chalmers, K. and Hockley, W. (2010) Item effects in recognition memory for words. *J. Mem. Lang.*, **62**, 1–18.
Gelman, A. and Hill, J. (2007) *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
Harrell, F. (2001) *Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.
Jaeger, T. (2008) Categorical data analysis: away from anovas (transformation or not) and towards logit mixed models. *J. Mem. Lang.*, **59**, 434–446.
Jara, A., Quintana, F. and San Martín, E. (2008) Linear mixed models with skew-elliptical distributions: a Bayesian approach. *Computnl Statist. Data Anal.*, **52**, 5033–5045.
Lewin, K. (1943) Defining the "field at a given time". *Psychol. Rev.*, **50**, 292–310.
McGue, M., Osler, M. and Christensen, K. (2010) Causal inference and observational research: the utility of twins. *Perspect. Psychol. Sci.*, **5**, 546–556.
Mukhopadhyay, S. and Gelfand, A. (1997) Dirichlet process mixed generalized linear models. *J. Am. Statist. Ass.*, **92**, 633–639.
Pinheiro, J. and Bates, D. (2000) *Mixed-effects Models in S and S-PLUS*. New York: Springer.
Quené, H. and Van den Bergh, H. (2008) Examples of mixed-effects modeling with crossed random effects and with binomial data. *J. Mem. Lang.*, **59**, 413–425.
Raudenbush, S. and Bryk, A. (2001) *Hierarchical Linear Models: Applications and Data Analysis Methods*. New York: Sage.
Ruts, W., De Deyne, S., Ameel, E., Vanpaemel, W., Verbeemen, T. and Storms, G. (2004) Dutch norm data for 13 semantic categories and 338 exemplars. *Behav. Res. Meth. Instr. Comput.*, **36**, 506–515.
Schwarz, G. (1978) Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.
Shi, Y., Leite, W. and Algina, J. (2010) The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *Br. J. Math. Statist. Psychol.*, **63**, 1–15.
Snijders, T. and Bosker, R. (2011) *Multilevel Analysis: an Introduction to Basic and Advanced Multilevel Modeling*, 2nd edn. London: Sage.
Stanley, J. (1961) Analysis of a doubly nested design. *Educ. Psychol. Measmnt*, **21**, 831–837.
Verbeke, G. and Molenberghs, G. (2000) *Linear Mixed Models for Longitudinal Data*. New York: Springer.
Wood, S. N. (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Statist. Soc.* B, **73**, 3–36.
Zeger, S., Liang, K. and Albert, P. (1988) Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.