DISCOURSE PROCESSES, *41*(3), 221–250

What Is Wrong With ANOVA and Multiple Regression? Analyzing Sentence Reading Times With Hierarchical Linear Models

Tobias Richter
University of Cologne, Germany

Most reading time studies using naturalistic texts yield data sets characterized by a multilevel structure: Sentences (*sentence level*) are nested within persons (*person level*). In contrast to analysis of variance and multiple regression techniques, hierarchical linear models take the multilevel structure of reading time data into account. They provide methods to estimate variance components and to model the influence of predictor variables on different levels as well as cross-level interactions between these predictors. This article gives a brief introduction to the method and proposes practical guidelines for its application to reading time data, including a discussion of power issues and the scaling of predictor variables. The basic principles of model building and hypothesis testing are illustrated with original data from a reading time study with naturalistic texts.

In research on text comprehension, reading times have been established as a useful and unobtrusive online measure to study resource allocation during reading. There are eye-tracking methods and the inexpensive, easy-to-use moving window method for recording such data (Haberlandt, 1994). For further statistical analysis, most researchers use analysis of variance (ANOVA) or multiple regression (MR) techniques. Neither of the two methods, however, represents an optimal way to deal with reading times because they cannot handle the typical multilevel structure of such data: Whenever a study requires a number of participants to read several sentences and reading times are recorded for each sentence, the resulting data set will have at least two levels, a person level and a sentence

Correspondence should be addressed to Tobias Richter, University of Cologne, Psychological Department, Herbert-Lewin-Str. 2, D-50931 Köln, Germany. E-mail: tobias.richter@uni-koeln.de

level. One important implication of this two-level structure is well known to psycholinguists since the seminal paper by Clark (1973). In reading time experiments, not only participants but also linguistic stimuli such as words or sentences are randomly sampled from larger populations. For this reason, inferential statistics must be based on two sources of error variance simultaneously if researchers wish to generalize their results across the particular stimuli used in a given study. However, the techniques of parameter estimation commonly used in ANOVA and MR are not designed to handle two sources of error variance at the same time.

Hierarchical linear models (Raudenbush & Bryk, 2002), also known as *random coefficient models* (Kreft & de Leeuw, 1998), *variance component models* (Longford, 1989), or *multilevel random coefficient models* (Nezlek, 2001, 2003), take the multilevel structure of reading time data into account. Hierarchical linear models have originally been developed in educational and social research where observations are often made on different levels simultaneously (e.g., students, classes, schools). Up to now, this type of method has not routinely been applied to reading time data despite its advantages: It is very well suited for typical research questions in the field, it provides a straightforward solution to methodological problems associated with the more traditional methods, and it offers new perspectives for research. A unique feature of hierarchical linear models is the possibility of estimating interaction effects of predictor variables located on different levels directly, for example interactions between sentence level and person level predictors.

The aim of this article is to give a short introduction to the application of hierarchical linear models to reading time data. A high emphasis is put on practical issues that are likely to be of primary interest to researchers in the field of reading and text comprehension. As a consequence, more technical aspects like procedures of parameter estimation are touched only briefly (for a more thorough coverage, see Goldstein, 2003; Raudenbush & Bryk, 2002).

This article starts with an explanation why and in which respects reading time data are structured by multiple levels. The second section then discusses methodological problems in the application of ANOVA and MR techniques to reading times characterized by a nested structure. Subsequently an outline of the principles of hierarchical linear models is given for a simple multilevel problem comprising of a sentence and a person level. Sample data from a reading time study are used to illustrate these principles and to demonstrate the advantages of hierarchical linear models compared to the more traditional models. The major part of the article covers basic issues of multilevel modeling. With a focus on reading time research, practical guidelines for model building and hypothesis testing are provided. Special emphasis is put on the issues of power of significance tests and scaling of predictor variables.

THE MULTILEVEL STRUCTURE OF READING TIME DATA

The main reason for applying hierarchical linear models to reading times is the typical multilevel structure of the data. In studies using naturalistic text material (Graesser, Magliano, & Haberlandt, 1994) as well as in experiments using short texts made up by the experimenter, participants read more than one sentence. Therefore, a simple two-level structure comprising a *person level* and a *sentence level* can be found in almost any reading time study. If several participants read several sentences and reading times are measured sentence by sentence, the variance in the sentence reading times will invariably have two sources: One proportion of variance will be due to differences between persons (variance on the person level), and a second proportion will be due to differences between sentences (variance on the sentence level). These two sources of variance will be present irrespective of the theoretical focus of a given study, be it set on predictors on the sentence or the person level, or on both levels simultaneously. Examples from research on text comprehension are of course abundant. In many experiments on inference processes, sentence characteristics are manipulated to investigate their influence on reading times (e.g., inference scores or causal relatedness, cf. Singer, 1994). The correlational approach of the three-pronged method employs think-aloud procedures to determine the probabilities for sentences to elicit specific inferences, and the results are used to predict sentence reading times (Magliano & Graesser, 1991). Predictors on the person level may be categorical variables such as experimentally manipulated factors (e.g., reading goals) as well as continuous variables depicting individual differences (e.g., reading skills). A number of studies entail comparisons of groups of readers (e.g., slow vs. fast readers, good vs. poor readers, younger vs. older readers; readers with different reading expectations, goals, or strategies) regarding how strongly sentence characteristics (e.g., microstructural vs. macrostructural features) affect reading times (e.g., Bisanz, Das, Varnhagen & Henderson, 1992; Graesser, Hoffman, & Clark, 1980; Haberlandt & Graesser, 1985; Long & Chong, 2001; Magliano, Trabasso, & Graesser, 1999; Zwaan, 1994). In terms of multilevel modeling, the focus of these studies is on cross-level interactions of predictors on the person and on the sentence level.

In text comprehension research, the stimulus materials themselves are often made up of different levels such as sentence, paragraph, and text. On each of these levels, there are various variables that might possibly influence reading times. In addition, the relationship between similar constructs may differ radically depending on the level of analysis (see Nezlek, 2001, for an illustration). The impact of syntactic complexity on reading times, for example, is probably much stronger when it is investigated on the sentence level (syntactic complexity and sentence reading times) compared to the text level (mean syntactic complexity and aggregated sentence reading times) because sentence-to-sentence variability of syntactic complexity contributes to text difficulty (Wisher, 1976). Generally stated, any reading time study warrants substantial conclusions only relative to one or more particular level(s) of analysis. If researchers disregard this principle, they run the risk to commit an ecological fallacy (Robinson, 1950).

Even if an investigator's theoretical perspective is restricted to one level, sources of variance on other levels will still be present and therefore must not be ignored. Accordingly, hierarchical linear models are often preferable to statistical methods that neglect the nested structure of reading time data.

WHAT IS WRONG WITH ANOVA AND MULTIPLE REGRESSION?

Typically, reading time studies employ ANOVA or MR techniques to investigate the impact of person- or sentence-level factors on the allocation of processing resources. Due to the fact that both types of methods are designed for single-level problems, the analysis of multilevel data such as reading times by ANOVA or MR is associated with serious problems.

ANOVA in its various forms is most prevalent in studies following an experimental approach. In reading time experiments, person-level characteristics such as reading goals, tasks, or reader expectations are usually varied as between-subjects variables, whereas characteristics of sentences, texts, or paragraphs are included as within-subjects variables. As a rule, dependent variables consist of aggregated measures. The rationale for using aggregated measures is to enhance reliability of dependent variables by eliminating variance caused by attributes specific to individual text segments (e.g., sentence length, passage difficulty, syntactic complexity). However, using person-level aggregates as dependent variables in ANOVA may be problematic because it results in an unnecessary loss of information and potential threats to the validity of the results. First of all, a large proportion of the variance between individual sentences may in principle be explained by easily available sentence attributes. In addition, interaction effects between attributes on the sentence level and attributes on the person level may exist. After aggregation, such cross-level interactions will remain unnoticed and the validity of person-level effects found in ANOVA models has to be questioned. Last but not least, as Clark (1973) pointed out (following a suggestion by Coleman, 1964), even if variance components due to sentences are regarded as pure error variance, researchers often commit a statistical fallacy if they exclude them from further analysis. To generalize effects across participants as well as sentences, hypothesis tests must be based on persons and sentences as sources of error; that is, also the effects of sentences must be treated as random rather than fixed.

In ANOVA, there is no way to estimate an error term that includes both sources simultaneously. The so called quasi *F* ratio *F*' proposed by Winer (1971) is available as an approximate test statistic for designs with subjects and sentences as random effects, but in many cases, for example in data sets with missing data, computation of *F*' is difficult or impossible. As a consequence, it has become customary in psycholinguistic papers to report two *F* tests, one based on persons as the source of error (F_1 based on the treatments by subjects interaction sums of squares) and one based on sentences as the source of error (F_2 based on the items within treatments sums of squares). Although the inflation of type-I-error probability is lower when two separate *F* tests are conducted, this procedure may still lead to false positive decisions because both *F* tests are biased when persons as well as sentences are sources of error (Raaijmakers, Schrijnemakers, & Gremmen, 1999). The less common alternative proposed by Clark (1973) is to base hypothesis tests on a statistic called *minF*' which is derived from F_1 and F_2. This procedure inflates type-II-error probability because *minF'* represents the lower bound of *F*' and may thus underestimate *F*'. In sum, even if the variance due to sentences is purely unsystematic, the hypothesis tests provided by ANOVA are not well designed to handle the multilevel structure of reading times. The common procedures that are used as workarounds yield either too progressive tests (inflating type-I-error probability) or too conservative tests (inflating type-II-error probability).

MR techniques have been introduced to reading time research to conduct more fine-grained analyses on the sentence level (cf. Graesser & Riha, 1984; Haberlandt, 1984). Nevertheless, MR models for sentence reading times face difficulties complementary to those in ANOVA, all of which reduce to the problem that sentence reading times come from different persons. Four methods to minimize the undesired effects of this problem have been proposed.

Sentence Reading Times as Independent Observations

One simple approach is to enter all sentence reading times (be they from the same or from different persons) into the regression model as independent observations. Some researchers following this approach *z* standardize the measures for each person to eliminate between-person variance. Regardless of the issue of standardization, treating all sentence reading times as independent observations amounts to ignoring the person level and may thus lead to completely erroneous conclusions. The strength of the association of a sentence-level predictor such as causal relatedness with reading times, for example, might vary between persons and depend on person characteristics such as the amount of prior knowledge. If cross-level interactions of sentence- and person-level predictors are present in the data set, they cannot be detected. Even worse, the MR model will be misspecified, and parameter estimates will be biased. Cronbach and Webb (1975) were the first to give a systematic analysis of this problem for the similar case of students nested within classes. In addition to biased estimates in the case of cross-level interactions, the traditional MR approach suffers from methodological deficiencies because it fails to separate variance components pertaining to different levels. Sentence reading times belonging to the same person, for example, usually show higher correlations than reading times from different persons, which means that the intraclass correlation of sentence reading times will be different from zero. One likely consequence of a high intraclass correlation is a violation of the homoscedasticity assumption, which is a precondition for conducting significance tests in MR models with ordinary least squares (OLS) estimates (e.g., Cohen, Cohen, West, & Aiken, 2003, chap. 4). In linear models with OLS estimates, a high intraclass correlation leads to an underestimation of the standard errors on which significance tests for individual parameters are based. Therefore, the actual alpha level will often be inflated compared to the nominal alpha level if the analysis neglects dependencies of reading times belonging to the same person (for the ANOVA case, see Barcikowski, 1981; for examples, see Kreft & de Leeuw, 1998).

Aggregation Across Persons

In a second, complementary approach, reading times are aggregated across persons, and mean (or sometimes median) reading times for each sentence are then used as the criterion variable in the regression equation. This approach faces similar problems as ANOVA because it disregards the person level. Because variance between persons is omitted by the aggregation procedure, the approach would be justified only if variances and covariances between persons were homogeneous across sentences—an assumption that is never tested and probably rarely met. In all other cases, parameter estimation may be biased if aggregated reading times are used. Moreover, aggregation across sentences again leads to an unnecessary loss of information, which can make the detection of cross-level interactions impossible.

Separate Regression Models

A third approach, which by now has become the most accepted approach in psycholinguistics and text comprehension research, makes use of a two-step procedure. The first step consists of estimating separate regression models for each person, with sentences as the units of analysis. In a second step, the parameters estimated for each person are compared to determine the interindividual consistency of estimates, or *t* tests are applied to test whether the mean coefficient across persons is significantly different from zero (cf. the first computational procedure proposed by Lorch & Myers, 1990). In some studies, an additional ANOVA or MR on the person level is conducted with the parameter estimates obtained for each person as dependent variable, to explain interindividual variability in these estimates (e.g., Juhasz & Rayner, 2003; Stine-Morrow, Millinder, Pullara, & Herman, 2001; Zwaan, 1994).

Clearly, this two-step procedure represents the closest approximation to a multilevel approach. Nevertheless, there are three interrelated arguments against the stepwise procedure. First, the total sample of reading time data is divided into many small subsamples (as many subsamples as there are participants in a given study), causing the reliability of parameter estimates to decrease (i.e., standard errors to increase) in comparison to estimates that exhaust the information contained in the total sample (Raudenbush & Bryk, 2002). Stated differently, the two step-approach requires that a high number of parameters (the number of parameters in each participant's submodel times the number of participants) must be estimated in step one. Given a finite set of data, a higher number of parameters to be estimated inevitably yields less reliable estimates. Aggregating the within-person coefficients in step two does not automatically compensate for the unreliability of the coefficients estimated in step one. The second and related argument against the two-step approach is that it does not account for interindividual variability in the reliabilities of parameter estimates (at least not if the standard OLS technique is used for parameter estimation). All parameter estimates from the first step are weighted equally in the analyses conducted in the second step, regardless of their standard error. For this reason, the results of the two-step approach may be biased. Here is a simple example: Imagine 2 participants who have each read the same three sentences. The reading times of Participant 1 are 3,000 ms, 3,100 ms, and 2,900 ms, whereas the reading times of Participant 2 are 1,000 ms, 8,000 ms, and 6,000 ms. Obviously, the 3,000-ms mean reading time of Participant 1 represents a more reliable estimate than the 5,000-ms mean reading time for Participant 2, which is reflected in differing standard errors for the two estimates. A two-step approach that employs OLS estimates, however, would treat the means of both participants as equally reliable and weigh both means equally when it comes to estimating the mean reading time across persons. A partial solution to this problem would be to use a technique called weighted least squares (WLS) for the estimation of the overall coefficients, which would take differences in the reliability of the individual parameter estimates into account (e.g., Cohen et al., 2003, chap. 4). But still, such an analysis would not be able to handle random error on both the sentence and the person level properly, which is the third argument against the two-step regression approach.

Generally speaking, the two-step approach fails to separate the variance components pertaining to sentence and person level in an appropriate way because these variances are considered sequentially, not simultaneously as in hierarchical linear models with random coefficients. In hierarchical linear models with random coefficients, the estimation of a large number of within-person coefficients is replaced by an estimation of parameters that describe their distribution, that is, the mean coefficient and its variance between persons (van der Leeden, 1998).

The estimation of parameters for individual persons is of course possible, but it is done in a way that attenuates for their unreliability (see the Parameter Estimation section).

Entering Persons as Dummy-Coded Variables in the Regression Equation

A fourth possibility is to enter persons as dummy-coded predictor variables into the regression model with sentence reading times as observational units (cf. the second computational procedure in Lorch & Myers, 1990). Entering persons as dummy-coded variables into the regression equation is a way to include the variance between the means of different persons in the model. This approach is often called least-squares dummy variable approach or fixed-effects approach to clustering because it treats the effects of higher level units as fixed, not as random (Snijders & Bosker, 1999, chap. 4.2). Beside this substantial difference to a hierarchical linear model with random coefficients, it would be quite cumbersome to model interindividual variability in sentence-level coefficients in the least-squares dummy variable approach. This is because the number of interaction terms necessary to meet this objective would explode soon, especially if designs with several independent variables are considered. Also, the coding system would have to be changed from a dummy-variable coding system to a coding system that yields centered variables to avoid nonessential multicollinearity of the person predictors and the product terms representing the interaction effects (Aiken & West, 1991; see also the Scaling and Coding of Predictor Variables section). The situation becomes even more complex when person-level predictors are to be included to explain interindividual variability of sentence-level coefficients. Accordingly, the least-squares dummy-variable approach is very rarely used in reading time research.

In sum, the application of both ANOVA and MR techniques to sentence reading times is likely to suffer from methodological problems and an unnecessary loss of information. These weaknesses result from reducing the multilevel structure inherent to reading time data to just one level of analysis. None of the solutions proposed to circumvent these weaknesses within the ANOVA and MR approaches are fully satisfactory. As will be shown in the next two sections, hierarchical linear models preserve the multilevel structure of reading time data. In contrast to ANOVA and MR techniques with OLS estimates, hierarchical linear models allow modeling error variance at both levels of analysis simultaneously. They not only avoid the methodological problems associated with the more traditional methods, but open up new and attractive perspectives for reading time research.

PRINCIPLES OF HIERARCHICAL LINEAR MODELING: A SIMPLE TWO-LEVEL EXAMPLE

This section illustrates the basic principles of hierarchical linear models in a simple two-level example with sentences as the lower level and persons as the higher level of analysis and with sentence reading times as criterion variable. The data come from a study where 38 psychology undergraduates read expository texts sentence-by-sentence; these were presented by a moving-window technique (self-paced, 77 sentences in total). Two independent variables, one located on the sentence level and one on the person level, were manipulated experimentally. For each participant, some of the sentences were modified in a way that they presented implausible information (*sentence plausibility*: plausible vs. implausible sentences). In addition, half of the participants were given the task of keeping in mind as much information from the texts as possible, whereas the other half were given the task of developing their own point of view regarding the validity of the information presented in the text (*reading goal*: memorization vs. standpoint goal).

Three hypotheses were tested: (a) It was expected that readers generally allocate more processing resources to sentences containing plausible information than to sentences containing implausible information, with the consequence that implausible sentences are read faster than plausible ones. Technically speaking, Hypothesis 1 predicts a main effect of the sentence-level predictor sentence plausibility. (b) It was also expected that readers who follow the goal of keeping in mind as much information as possible process the text more thoroughly than readers who read to develop their own point of view, resulting in longer reading times for the memorization instruction. Accordingly, Hypothesis 2 predicts a main effect of the person-level predictor reading goal. (c) Readers who process the text with the goal of developing their own standpoint, however, should devote additional resources to strategic evaluative processes when they encounter implausible sentences. For this reason, the effect predicted by Hypothesis 1 should be weakened by the standpoint instruction. Thus, Hypothesis 3 predicts an ordinal cross-level interaction of the person-level predictor reading goal and the sentence-level predictor sentence plausibility. These hypotheses illustrate three basic types of effects, which may be tested in multilevel models of sentence reading times. The notation used in the example and the following parts of the article is the one proposed by Raudenbush and Bryk (2002).

A good point to start with building a hierarchical linear model is always to set up the simplest of all possible multilevel models, a so-called *unconditional model*. An unconditional model does not contain any predictors but only an intercept term plus an error term for the lowest level of analysis ("level 1" in Raudenbush & Bryk, 2002; "individual level" in Kreft & de Leeuw, 1998), and an error term for the next highest level of analysis ("level 2" in Raudenbush &

Bryk, 2002; "contextual level" in Kreft & de Leeuw, 1998). The unconditional model may be written in two parts:

$$Y_{\mathrm{ij}} = \beta_{0\mathrm{j}} + r_{\mathrm{ij}}, \tag{1a}$$

$$\beta_{0\mathrm{j}} = \gamma_{00} + u_{0\mathrm{j}}. \tag{1b}$$

In Equation 1a, which is the sentence-level part of the unconditional model, the double-indexed criterion variable Y_{ij} represents sentence reading times, with the index i denoting sentences and the index j denoting persons. Accordingly, the intercept $\beta_{0\mathrm{j}}$ is the estimated mean reading time for each person, and r_{ij} is the sentence-level error term. Equation 1b, which is the person-level part of the unconditional model, models the sentence-level intercept as a function of the grand-mean intercept γ_{00} and a person-level error term $u_{0\mathrm{j}}$. The presence of two error terms, one assigned to the sentence level and one for the person level, marks a distinctive feature of hierarchical linear models with random coefficients and a major difference from traditional regression models with OLS estimation. In contrast to such models, where both variance components cannot be separated, the variances of error terms on both levels may be estimated in hierarchical linear models. Later on in this section, the meaning of these variance components will be discussed in more detail. Because the unconditional model does not contain any predictors, the variances of $u_{0\mathrm{j}}$ and r_{ij} together capture all of the criterion variance. The sentence-level variance component σ^2 is based on the deviances of particular sentence reading times from their respective person mean $\beta_{0\mathrm{j}}$ and the person-level component τ_{00} is based on the deviances of particular person means from the grand mean γ_{00}. For the sample data, the variance σ^2 of the sentence-level error term in the unconditional model is estimated as 86416677, and the variance τ_{00} of the person-level error term is estimated as 20098075 (the variances are so large because reading times were measured in milliseconds). Using these variances, we can compute the *intraclass correlation coefficient* ρ:

$$\begin{aligned}\rho &= \tau_{00} / (\tau_{00} + \sigma^2) \\ &= 20098075/(20098075 + 86416677) \\ &= .19.\end{aligned} \tag{2}$$

The intraclass correlation coefficient is defined as the proportion of criterion variance between level-2 units; it is identical to the η^2 measure, which is used in ANOVA to denote effect sizes. In a model for sentence reading times, the intraclass coefficient represents the proportion of variance due to differences between persons in relation to the overall variance in sentence reading times. The intraclass coefficient of .19, which we have obtained for the sample data, indicates that a considerable proportion of variance pertains to the sentence as well as the person level. Only if the intraclass coefficient is zero or close to zero can one conclude that there is no criterion variance on level 2. Even in this (highly unlikely) scenario, it would make sense to proceed with building a conditional multilevel model, that is, a multilevel model that contains predictors, instead of a one-level regression model. This is because the slopes of any sentence-level predictors, which are included in the model, might vary randomly or as a function of person-level predictors.

For the sample problem, we continue with a simple conditional model for the sentence level. This model includes the continuous predictor *number of syllables* to control for different lengths of sentences, represented by X_1, and the dichotomous predictor *sentence plausibility*, represented by X_2, which is relevant for Hypotheses 1 and 3. To facilitate interpretation of coefficients, we include number of syllables as grand-mean centered predictor, which means that it is centered around the mean number of syllables across all sentences and persons. In this case, grand-mean centering is equivalent to group-mean centering, that is, centering the predictor around its person means, because the total number of syllables in the texts was held constant for all participants (for a general discussion of centering and coding options, see the Scaling and Coding of Predictor Variables section). Sentence plausibility is coded with weighted-effects coding, a coding technique that yields centered variables. Thus, the conditional model for the sentence level is

$$Y_{\mathrm{ij}} = \beta_{0\mathrm{j}} + \beta_{1\mathrm{j}}(X_{1\mathrm{ij}} - \overline{X}_1) + \beta_{2\mathrm{j}} X_{2\mathrm{ij}} + r_{\mathrm{ij}}. \tag{3}$$

In Equation 3, the regression coefficient $\beta_{1\mathrm{j}}$ indicates the magnitude and direction of the relationship between number of syllables and the sentence reading times. Due to the centering and coding options chosen for the two predictor variables, the sentence-level intercept $\beta_{0\mathrm{j}}$ denotes an estimate of the mean reading time for sentences with an average number of syllables, with plausible and implausible sentences contributing equally to the estimate. Finally, the sentence-level model contains the error term r_{ij}. So far, the sentence-level model resembles an ordinary regression model, except for one small but important detail: Just as in the unconditional model, all variables as well as the error term carry two indexes, index i, which denotes one particular sentence, and index j, which denotes one particular person. Both the sentence-level intercept $\beta_{0\mathrm{j}}$ and the slopes $\beta_{1\mathrm{j}}$ and $\beta_{2\mathrm{j}}$ are indexed by the letter j, which means that both sentence-level parameters are allowed to vary between persons as the observational units on the higher level.

Consequently, the next step is to construct a person-level model for each of the three sentence-level parameters:

$$\beta_{0\mathrm{j}} = \gamma_{00} + \gamma_{01} W_{\mathrm{j}} + u_{0\mathrm{j}} \tag{4a}$$

$$\beta_{1\mathrm{j}} = \gamma_{10} + \gamma_{11} W_{\mathrm{j}} + u_{1\mathrm{j}}, \tag{4b}$$

$$\beta_{2\mathrm{j}} = \gamma_{20} + \gamma_{21} W_{\mathrm{j}} + u_{2\mathrm{j}}. \tag{4c}$$

In the *intercept model* in Equation 4a, the sentence-level intercept β_{0j} for each person j is predicted by the person-level intercept γ_{00} and reading goal as the person-level predictor W_j associated with the slope γ_{01}, and a person-specific error term u_{0j}. Reading goal is included as a contrast-coded predictor (–1 for the memorization goal, 1 for the standpoint goal), a coding option that yields centered variables when the sizes of the coded groups are equal (for a general discussion, see the Scaling and Coding of Predictor Variables section). Due to the coding and scaling options chosen for reading goal and the two sentence-level predictors, the person-level intercept γ_{00} represents an estimate of the weighted mean of reading time for plausible and implausible sentences with an average number of syllables. The slope γ_{01} represents the direction and magnitude of the departure of the reading goal conditions from the overall mean, that is, the main effect of the reading goal manipulation that is relevant for Hypothesis 1. Finally, Equation 4a contains an error component u_{0j}, which represents random, between-person fluctuation in the sentence-level intercepts. The submodels in Equation 4b and 4c, which are the models for the sentence-level slopes of number of syllables and sentence plausibility, include the person-level intercept γ_{10} and γ_{20}, respectively, which represent the mean effect of these sentence-level predictors, the person-level slopes γ_{11} and γ_{21}, which represent the slope of the person-level predictor reading goal, plus the error terms u_{1j} and u_{1j}, which capture random, between-person fluctuations of the effects of number of syllables and sentence plausibility between persons. Both submodels are *slope-as-outcome models* because a portion of the interindividual variability in the sentence-level slopes is to be explained by the person-level predictor reading goal. Reading goal is again included as a contrast-coded predictor. As a consequence of the coding and centering options chosen, the intercepts γ_{20} in Equation 4c represents the mean effect of sentence plausibility on sentence reading times, that is, the main effect of sentence plausibility that is relevant for Hypothesis 2. The slope γ_{21} describes the systematic variation of the effect of sentence plausibility as a function of reading goal. It represents the cross-level interaction of reading goal and sentence plausibility, which is critical for Hypothesis 3. The variance of the error component u_{2j} captures additional random fluctuation in β_{2j}.

The sentence-level model in Equation 3 and the three person-level models in Equations 4a, 4b, and 4c may be written into one equation, the *combined model*:

$$
\begin{aligned}
Y_{ij} &= \gamma_{00} + \gamma_{01} W_j + \gamma_{10}(X_{1ij} - \overline{X}_1) + \gamma_{20} X_{2ij} + \gamma_{11} W_j (X_{1ij} - \overline{X}_1) + \gamma_{21} W_j X_{2\,ij} \quad \textit{fixed part} \\
&\quad + u_{0j} + u_{1j}(X_{1ij} - \overline{X}_1) + u_{2j} X_{2ij} + r_{ij}. \qquad \textit{random part} \qquad (5)
\end{aligned}
$$

The error terms u_{0j}, $u_{1j}(X_{1ij} - \overline{X}_1)$, and $u_{2j}X_{2j}$ in the random part of the combined model represent a distinctive feature of hierarchical linear models. Because these error terms were included in Equations 4a, 4b, and 4c the sentence-level intercept β_{0j} as well as the sentence-level slopes β_{1j} and β_{2j} are modeled as *random coeffi-*

cients. Random coefficients are assumed to have a random fluctuation between observational units on the next highest level of analysis, in this case on the person level. The random fluctuation of the sentence-level intercepts is captured in the variance of the person-level error term u_{0j}, which is denoted by τ_{00}. Accordingly, the random fluctuation of the person-level slopes for number of syllables and sentence plausibility are captured in the variance of the person-level error terms u_{1j} and u_{2j}, which are denoted by τ_{11} and τ_{22}, respectively. The person-level error terms are also allowed to covary. These covariances are designated with τ_{01}, τ_{02}, and τ_{12}. If, for example, persons who read more slowly (over and above what is predicted by the sentence-level predictor reading goal) were also affected more strongly by the number of syllables in a sentence, u_{0j} and u_{1j} would exhibit a positive covariation. The variances and covariances of the person-level error terms and the variance σ^2 of the sentence-level error term r_{ij} are called *variance components*.

In contrast to random coefficients, *fixed coefficients* are assumed to vary systematically across units on the higher level of analysis, and no random error term is estimated. The choice to model coefficients as fixed or random distinguishes hierarchical linear models from more traditional linear models with OLS estimates, where only fixed coefficients may be included. Traditional linear models rest on the presupposition that there is no error variance between observational units on any but the lowest level of analysis (otherwise, the homoscedasticity assumption of single-level models would be violated). The more complex error structure of hierarchical linear models, in contrast, allows the inclusion and estimation of variance components on each level of analysis. Estimation of variance components in models with random coefficients does not work with the OLS method. Instead, ML techniques are most commonly used.

For the sample data, we use an estimation technique called restricted maximum likelihood in combination with generalized least squares estimates (see the Parameter Estimation section). For the significance tests, we use single-parameter *t* tests that are based on the ratio of the parameter and its standard error (see the Hypothesis Testing section). These tests correspond to those that may be used to test single parameters in ordinary regression models. The parameter γ_{01}, which represents the main effect of reading goal, is estimated as –1422, with a standard error of 689, and it is significantly different from zero, $t(36) = -2.1$, $p < .05$ (Table 1). In line with Hypothesis 1, the standpoint goal lead to generally faster reading compared to the average reading time, whereas the memorization goal lead to slower reading. The parameter γ_{10}, which represents the main effect for number of syllables, is estimated as 245 ($SE = 16$), $t(2920) = 15.2$, $p < .001$. Not surprisingly, the more syllables a sentence contained, the more time it took participants to read it. The parameter γ_{20}, which represents the main effect for sentence plausibility, is estimated as –822 ($SE = 232$), $t(2920) = -3.5$, $p < .01$. In line with Hypothesis 2, implausible sentences were generally read faster, whereas plausible sentences were read slower. However, the parameter γ_{21}, which represents the cross-level interaction of

TABLE 1
Estimates of the Fixed Effects From a Hierarchical Linear Models Analysis of the Sample Data Compared to Estimates From a One-Level Multiple Regression Analysis and a Two-Step Separate Regressions Analysis

	Hierarchical Linear Model[a]			*One-Level Multiple Regression*			*Two-Step Separate Regressions*		
Fixed Effect	*Coefficient*	*SE*	*t(df)*	*Coefficient*	*SE*	*t(2920)*	*Coefficient*	*SE*	*t(37)*
Intercept γ_{00}	12740	689	18.5***(36)	12741	169	75.2***	12829	752	17.1***
Reading goal γ_{01}	–1422	689	–2.1*(36)	–1420	169	–8.4***	–1443	725	–2.0
Number of syllables γ_{10}	245	16	15.2***(2920)	246	9	26.4***	245	17	14.6***
Number of Syllables × Reading Goal γ_{11}	–20	16	–1.2(2920)	–19	9	–2.0*	–21	17	–1.2
Sentence plausibility γ_{20}	–822	232	–3.5**(2920)	–911	266	–3.4**	–830	260	–3.2**
Sentence Plausibility × Reading Goal γ_{21}	501	232	2.2*(2920)	350	263	1.3	483	252	1.9

Note. Reading goal (contrast coding, centered): memorization goal (–1) versus standpoint goal (1); number of syllables: grand-mean centered; sentence plausibility (weighted effects coding, centered): plausible sentences (–.42) versus implausible sentences (1).

[a] Restricted Maximum Likelihood/Generalized Least Squares estimates.

**p* < .05. ***p* < .01. ****p* < .001 (two-tailed).

sentence plausibility and reading goal, is estimated as 501 (*SE* = 232) and it is also significant, $t(2920) = 2.2$, $p < .05$. Thus, in line with Hypothesis 3, when they read implausible sentences, participants following the standpoint goal did not speed up as much as participants following the memorization goal. For interpreting this interaction, it is helpful to use the parameter estimates to generate predicted values (i.e., estimated cell means) for different combinations of the interacting variables. For ease of interpretation, we assume that number of syllables takes on a mean value and its interaction with reading goal is zero, so that we can ignore this predictor in the predictions. The following values are predicted:

1. Plausible sentence, memorization goal:
   $12{,}740 + (–1{,}422 \times –1) + (–822 \times –0.42) + (501 \times –0.42 \times –1) = 14{,}717.66$
2. Plausible sentence, standpoint goal:
   $12{,}740 + (–1{,}422 \times 1) + (–822 \times -0.42) + (501 \times –0.42 \times 1) = 11{,}452.82$
3. Implausible sentence, memorization goal:
   $12{,}740 + (–1{,}422 \times –1) + (–822 \times 1) + (501 \times 1 \times –1) = 12{,}839.00$
4. Implausible sentence, standpoint goal:
   $12{,}740 + (–1{,}422 \times 1) + (–822 \times 1) + (501 \times 1 \times 1) = 10{,}997$

The cross-level interaction is illustrated by the fact that the difference between memorization goal and standpoint goal is much larger in plausible sentences (3,265 ms) than it is in implausible sentences (1,842 ms).

The estimates for the variances of the person-level error terms are 18845487 for τ_{00}, 8722 for τ_{11}, and 253745 for τ_{22}. Significance tests of these variance components are based on the χ^2 statistic (see the Hypothesis Testing section). In this case, they reveal that only τ_{00}, the variance component of the intercept model, and τ_{11}, the variance component of the model for the slope of number of syllables, are significantly different from zero, whereas τ_{22}, the variance component of sentence plausibility, is not. Substantially, this means that there is considerable random variation between persons for the intercept as well as for the effect that number of syllables has on reading times. For the effect of sentence plausibility, however, systematic and random variation cannot be separated reliably. For this reason, we may fix the variance component τ_{22} and drop the error term u_{2j} from the model. The random coefficient for sentence plausibility then becomes a fixed coefficient.

Some points are worth noting here that might help to elucidate the relationship of the sample hierarchical linear model to traditional linear models. The first point is that the person-level variance components we have just estimated do not reduce to between-person variance estimates as they are used to construct the error terms for the *F* tests in a traditional (fixed-effects) ANOVA. Rather, they represent random fluctuation in the sentence-level intercept and sentence-level coefficients. The person-level residuals are conceived of as randomly drawn from a larger population and the sentence-level intercept and coefficients associated with a person-level error term are regarded as random coefficients. Note that the random part of the model in Equation 5 would reduce to the sentence-level error term r_{ij} if all person-level variance components were fixed to zero. In this case, the model would still be a multilevel model (and a direct extension of the separate regressions approach mentioned in the previous section), but no longer a random coefficient model. Instead, it would become a model with nonrandomly varying coefficients, which could be estimated by standard OLS techniques (for a systematic account of this type of model, see Burstein, Linn, & Capell, 1978). For the sample data, however, fixing the significant variance components to zero would introduce severe bias into the results, as will be illustrated in the following section.

COMPARISON OF THE SAMPLE HIERARCHICAL LINEAR MODEL TO OTHER METHODS

The previous sections argued that hierarchical linear models are often preferable to traditional linear models such as ANOVA and MR. There are cases, of course, in which these methods will lead to similar conclusions, but for the majority of cases—whenever the variance in reading times is distributed over more than one level—they may produce more or less different results. Especially when effects have to be modeled as random, the results of a hierarchical linear models analysis may be expected to differ from those of the more traditional methods that are restricted to modeling fixed effects. This section exemplifies this point by comparing the results of the sample hierarchical linear model devised in the previous section to the results of three well-established competitors: (a) a mixed-factors ANOVA based on reading times aggregated across sentences, (b) an MR analysis with all sentence reading times treated as independent observations, and (c) an MR analysis based on separate regression models for different persons.

An ANOVA based on reading times that are aggregated across sentences requires controlling for sentence length, that is, number of syllables. The most adequate approach advocated in the literature involves conducting separate simple regressions for each person with the number of syllables as predictor, and then aggregating the residuals from these regression analyses. In a mixed-factors ANOVA based on these residuals, only the strong main effect for sentence plausibility was significant, $F_1(1, 36) = 13.1, p < .01, \eta^2 = .27$. Plausible sentences were read slower ($M = 458$, $SE = 94$) than implausible sentences ($M = –773$, $SE = 252$). Neither the main effect of reading goal, $F_1(1, 36) = 1.7, p > .05$, nor the interaction of sentence plausibility and reading goal were significant, $F_1(1, 36) = 3.8, p > .05$. Thus, two of the effects that were significant in the hierarchical linear models analysis were not detected by ANOVA. There are two likely and related causes for this relative lack of power. The first one is the loss of information due to the aggregation procedure, and the second one is the failure of ANOVA to separate variance pertaining to the person and the sentence level in an appropriate way.

For the MR analysis with OLS estimates and sentence reading times treated as independent observations, all predictor variables were coded in the same way as in the hierarchical linear models analysis and were entered simultaneously into the model to allow direct comparisons with the parameter estimates of the full hierarchical linear model devised in the previous section (Table 1). There was a strong main effect for number of syllables—$b = 246$, $SE = 9$, $t(2920) = 26.4$, $p < .001$—and weaker but significant main effects for reading goal—$b = –1420$, $SE = 169$, $t(2920) = –8.4$, $p < .001$—and sentence plausibility—$b = –911$, $SE = 266$, $t(2920) = -3.4$, $p < .01$. The interaction term of sentence plausibility and reading goal, which was significant in the hierarchical linear models analysis, failed to reach significance in the MR analysis—$b = 350$, $SE = 263$, $t(2920) = 1.3$, $p > .05$. Instead, the unexpected interaction of reading goal and number of syllables was significant—$b = –19$, $SE = 9$, $t(2920) = –2.0$, $p < .05$. Compared to the hierarchical linear models analysis, the sentence-level MR analysis overestimated the slope for the main effect of sentence plausibility and underestimated the interaction effect of sentence plausibility and reading goal. Moreover, the standard errors of four parameters, the intercept, and the main effects for reading goal and number of syllables as well as the interaction of reading goal and number of syllables, were noticeably underestimated, in one case leading to a false positive error. These differences result from the less complex error structure of the sentence-level MR analysis. The two random effects that were significant in the full hierarchical linear model cannot be included in a single-level MR analysis with OLS estimates, with the inevitable consequence of a misspecified model.

For the MR approach based on separate regression analyses, all predictors were again coded in the same way as in the hierarchical linear models analysis to allow direct comparisons (Table 1). In a first step, separate regression analyses were conducted for each participant, with number of syllables and sentence plausibility entered simultaneously. The average intercept was 12829 ($SE = 725$), the average slope for number of syllables was 245 ($SE = 17$), and the average slope for sentence plausibility was –830 ($SE = 260$). All three coefficients were significantly different from zero—for all tests: $|t|(37) > 3.1$, $p < .01$. Thus, the separate regressions approach indicated main effects for both sentence-level predictors. In the next step, person-level models were employed to test for the main effect of reading goal (with the sentence-level intercept as criterion variable) and the interaction effect of reading goal with number of syllables and sentence plausibility (with the slope of sentence plausibility as criterion variable). In these analyses, neither the main effect of reading goal ($b = –1443$, $SE = 725$) nor the interaction effect with sentence plausibility ($b = 448$, $SE = 252$) were significant—for both tests: $|t|(37) < 2.0$, $p > .05$. Thus, although the separate regressions approach may be regarded as the strongest competitor of a hierarchical linear models analysis, it failed to detect two of the effects that were significant in the hierarchical linear models analysis. Table 1 shows that the two-step analysis slightly overestimates the slopes. The standard errors, however, are considerably overestimated, too, which may lead to false negative errors in hypothesis tests. From a methodological as well as an epistemological perspective, falsely concluding that a hypothesis is not valid may be just as harmful as falsely accepting a hypothesis as being valid. Moreover, it is always desirable to use the most precise data analysis methods available. For these reasons, the sample data strongly underscore the demand for using multilevel analyses analyzing reading times.

HIERARCHICAL LINEAR MODELS FOR READING TIME DATA: BASIC ISSUES

This part of the article attempts to provide a more detailed idea of how the application of hierarchical linear models to reading times works in practice. A number of basic issues central to every study that involves multilevel models will be discussed. These issues include a general strategy for exploratory model building, a description of the assumptions that characterize a well-specified model, the important aspects of scaling and coding of predictor variables, parameter estimation, hypothesis testing, and the relation of sample sizes and power. The following comments are intended to give reading time researchers practical guidelines and clues for orientation in this rather complex field. Therefore, they represent a practical advance organizer rather than a substitute for more general introductions into the topic.

Model Building and Assumptions Underlying Hierarchical Linear Models

Hierarchical linear models provide very flexible means for model building. One part of this flexibility stems from the opportunity to introduce several predictor variables on each level and from the option to build models which encompass more than two levels—provided that the data set is structured in an appropriate way and that it contains enough information (see the Power and Required Sample Sizes section). For the two-level case, the general formulation of a level-1 model (sentence-level model, cf. Equation 3) with Q predictor variables is

$$Y_{\mathrm{ij}} = \beta_{0\mathrm{j}} + \beta_{1\mathrm{j}} X_{1\mathrm{ij}} + \beta_{2\mathrm{j}} X_{2\mathrm{ij}} + \mathrm{K} + \beta_{\mathrm{Qj}} X_{\mathrm{Qij}} + r_{\mathrm{ij}}. \tag{6}$$

The corresponding general level-2 models (person-level models, cf. Equation 4) with a maximum of S predictor variables each are

$$
\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01} W_{1j} + \gamma_{02} W_{2j} + \mathrm{K} + \gamma_{0S} W_{Sj} + u_{0j}, \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} W_{1j} + \gamma_{12} W_{2j} + \mathrm{K} + \gamma_{1S} W_{Sj} + u_{1j}, \\
\beta_{2j} &= \gamma_{20} + \gamma_{21} W_{1j} + \gamma_{22} W_{2j} + \mathrm{K} + \gamma_{2S} W_{Sj} + u_{2j}, \\
\beta_{Qj} &= \gamma_{Q0} + \gamma_{Q1} W_{1j} + \gamma_{Q2} W_{2j} + \mathrm{K} + \gamma_{QS} W_{Sj} + u_{Qj}.
\end{aligned} \tag{7}
$$

Note that the level-1 model in Equation 6 as well as the level-2 models in Equation 7 may include main effect terms and within-level interaction terms. Furthermore, the predictors in models on both levels may be scaled in different ways, which may change the meaning of intercepts and slopes (see the Scaling and Coding of Predictor Variables section). The general model can be adapted very flexibly by fixing parameters in particular submodels to zero. The person-level models, for example, can differ in the set of person-level predictor variables (although it is customary to include the same higher level predictors in all submodels). Moreover, the inclusion of an error term u_{qj} is optional. In case this error term is omitted the respective coefficient β_{qj} turns from a random coefficient into a fixed coefficient.

The high degree of flexibility also implies that researchers have to face a number of decisions: Which variables should be included in the model(s), which coefficients should be modeled as fixed or random, and when should within-level interactions or cross-level interactions be modeled? Starting from sound theoretical assumptions is always important, but most model building will rely on exploration to a certain degree. Hox (1995) has proposed a stepwise strategy for exploratory model building that starts with an unconditional model as introduced here in the context of the sample hierarchical linear model. Following Hox (1995), the next steps would consist of setting up a sentence-level model, first without and then with variance components (random effects). Finally, person-level predictors, including cross-level interactions, may be included. Ranging from the simplest possible model to ever more complex models, these models form a sequence of nested models. At each step, the improvement of model fit achieved by adding more elements to the model may be tested for significance (see the Hypothesis Testing section). The result of all model building efforts—be they exploratory or theory driven—should be a well-specified model, which fulfills some general assumptions concerning the residuals on different levels. These assumptions are briefly outlined in the next section.

General Assumptions Underlying the Application of Hierarchical Linear Models

Similar to single-level linear models with OLS estimates, the application of hierarchical linear models rests on a number of basic assumptions. Violation of these assumptions leads to a misspecification of the model. As a consequence, parameter estimates and their standard errors may be biased, which also renders significance tests invalid (see Raudenbush & Bryk, 2002, chap. 9). The assumptions underlying hierarchical linear models refer to the distributions of the error components as well as to their relationships to each other and to the predictor variables. For the level-1 error component r_{ij}, normal distribution and a constant variance within each level-2 unit (homoscedasticity) are required. Similarly, the joint distribution of the level-2 error components u_{qj} is assumed to be multivariate normal with a constant covariance matrix across level-2 units. In addition, all error components should be independent of the predictor variables and have population means of zero.

In principle, several factors may cause violations of these assumptions. Important explanatory variables might be disregarded in the model, or a randomly varying slope could be erroneously specified as a fixed effect. In addition, a cross-level interaction or a within-level interaction, which is in fact present, might be missing from the model. Some of the predictor variables (e.g., interindividual difference measures on the person level) might be associated with a large amount of measurement error. Furthermore, there might be anomalies in the reading time data, such as outliers. Another possible source for misspecification is the fact that distributions of reading times are typically highly skewed to the right, with a peak in the lower range and a long tail.

The best way to avoid problems with misspecification is to design experiments carefully, to check the data for outliers and other anomalies, and to guide model building by a thorough step-by-step strategy starting from the sentence level. A transformation, which makes the reading time distribution more symmetric, is often an effective remedy for problems with the distributions of residuals. Both reciprocal and logarithmic transformations (Ratcliff, 1993), for example, may reduce biasing influences of outliers and help to avoid misspecification due to nonnormality of residuals. In addition to these general recommendations, Snijders and Bosker (1999) and Snijders (in press) proposed several checks for possible violations of model assumptions. At this point, it should also be noted that the logic of hierarchical models can in principle be extended to the analysis of nonmetric criterion variables such as binary or multicategory data (Snijders & Bosker, 1999, chap. 14; Wong & Mason, 1985). These types of analysis rest on assumptions different from those on which hierarchical linear models are based.

Scaling and Coding of Predictor Variables

In multilevel models, scaling of predictor variables is of much greater importance than in MR models that contain main effects only. The scaling of predictors on lower levels has a huge impact on the interpretation, the estimation, and significance tests of parameters on the same and higher levels. This section first sketches basic scaling options for continuous variables. Subsequently, the effects of different ways of coding categorical variables are briefly discussed.

Scaling of continuous variables. For continuous predictors on lower levels, centering around the mean of a higher level unit ("group-mean centering") and centering around the mean of the total sample ("grand-mean centering") are the most prevalent scaling options (Kreft, de Leeuw, & Aiken, 1995). Centering, that is, transforming the original predictor variables into deviation scores by subtracting a mean value, is especially useful for altering the meaning of the intercept term. In models with raw (i.e., noncentered) scores as predictor variables, the lower level intercept corresponds to the expected value of the criterion Y_{ij} when all predictors take on the value zero. For many predictor variables (in reading time research and elsewhere in psychology), however, the value zero can take on a problematic meaning. It does not make sense, for example, to estimate the mean reading time of sentences without any syllables.

Group-mean centering is one way to solve this problem. If every level-1 predictor is centered in this way, the level-1 intercept will represent the expected criterion values for the level-2 units when all level-1 predictors take on mean values relative to the respective level-2 unit. This form of centering is useful if there are theoretical reasons to omit predictor variance between level-2 units from the model. There is, however, the possibility of reintroducing this variance by entering the respective means of the level-1 predictors as predictor variables on level-2. Consider, for example, participants in a nonexperimental or not perfectly balanced study who have not read identical text materials. In a study like this, there is variance in sentence-level predictors like the number of syllables between sentences as well as between persons. Using group-mean centering for number of syllables omits the between-person portion of this variance on the sentence level, but we can reintroduce it into the model by using the mean number of syllables as a person-level predictor. Nevertheless it is important to keep in mind that centering of predictors around the mean of higher level units usually does not yield models that are statistically equivalent to models with the raw scores as predictors (for a discussion see Kreft et al., 1995).

The situation is different for the second common centering option, *grand-mean centering*. This technique is useful especially if interactions between predictors located on the same level are included because nonessential multicollinearity of main effect and interaction terms can be avoided (see Aiken & West, 1991). If every level-1 predictor is centered around the grand mean, the level-1 intercept will represent the expected value of the level-2 units when all level-1 predictors take on the mean of the total sample of reading times (*adjusted means*). Grand-mean centering yields models that are statistically equivalent to a raw score model regarding model fit, residuals, and predicted values (Kreft et al., 1995). Still, both models will produce different estimates for the intercepts and, consequently, estimates as well as significance tests for parameters in the intercept model will differ.

In experimental reading time studies with standardized text material (identical or strictly parallel texts for each participant), there should not be any between-person variance in sentence-level predictors (cf., e.g., the sample hierarchical linear model introduced in the Principles of Hierarchical Linear Modeling section). In this case, group-mean centering and grand-mean centering are equivalent options. But even for experimental studies, centering around the person mean (with reintroducing the between-person variance on the person level) can be a reasonable option, whenever experimental conditions are not based on the same text material for every participant. This is the case, for example, if texts and experimental conditions are combined in an incomplete, unbalanced design. In this case, centering sentence-level predictors like number of syllables around the respective person means, and reintroducing the person means as person-level predictors, provides an appropriate way of controlling for differences in the texts that otherwise contribute to error variance.

Coding of categorical predictors. To investigate the influence of categorical variables in regression models, whether measured or manipulated experimentally, a coding scheme has to be used, with *k*-1 separate predictors for a categorical variable with *k* categories. Each of the coding methods that have been proposed in the context of traditional linear models (see, e.g., Cohen et al., 2003, chap. 8) is also applicable to hierarchical linear models. Again, the major considerations in the choice of a coding scheme refer to the meaning of parameters at the various levels. Among the most common methods are dummy coding, contrast coding, and effects coding.

In *dummy coding*, one category serves as a reference category and is assigned the value 0 in all code variables. As a consequence, the intercept of a sentence-level model with a dummy-coded categorical variable (and no other predictors) will represent the expected value of reading times for sentences from the reference category, and the level-1 regression coefficients express the deviations of sentences belonging to one of the other categories from sentences belonging to the reference category. *Contrast coding* involves the construction of centered and orthogonal coding variables, which provide flexible means to test a priori hypotheses about differences between categories (e.g., levels of an experimental factor such as the reading goal manipulation in the sample hierarchical linear model). The intercept in models with contrast coded variables contains the unweighted means across all observational units. *Effects coding* is chosen if differences of categories from the overall mean of the criterion variable are of interest. There are two different variants of effects coding, *unweighted* and *weighted* effects coding. In weighted effects coding, the code variables are constructed in a way that reflects the relative category sizes (e.g., different proportions of sentence types as in the sample hierarchical linear model). As a consequence, weighted effects codes are centered variables. In a sentence-level model with weighted effects codes, the intercept will represent the weighted mean of all sentences, and the regression coefficients for the code variables will express deviations of sentences belonging to one sentence type from the weighted mean.

Throughout the discussion of coding options and their effects on the interpretation of intercepts and coefficients, it has been assumed that the codes are entered into the model in the way they have been constructed, which means as centered predictors when a coding option yields centered variables (such as in weighted effects coding), or as uncentered predictors when a coding option yields uncentered variables (such as dummy coding). This procedure, which has also been applied in the sample hierarchical linear model (see the Principles of Hierarchical Linear Modeling section), is most common because it leaves the standard interpretation of coefficients intact.

Parameter Estimation

In hierarchical linear models, parameters are estimated simultaneously on all levels of analysis. In addition to intercept and slope coefficients, variance components (variances as well as covariances of higher level error terms) have to be estimated when the model contains random coefficients. With OLS techniques, it is impossible to estimate these variance components. Instead, maximum likelihood (ML) techniques in combination with so called empirical Bayes estimates and an iterative algorithm (Expectation-Maximization algorithm) are commonly used (Goldstein, 2003; Raudenbush & Bryk, 2002). The principle of empirical Bayes-estimates implies that lower level parameters (e.g., the sentence-level coefficients), which belong to different higher level units (e.g., different persons), are weighted by their reliability. The weighing procedure corrects estimates from small or heterogeneous subsamples in the direction of the average estimates for the total sample (shrinkage; for illustrations see Kreft & de Leeuw, 1998). For example, parameter estimates for a participant of whom reading times from only few sentences are available, or participants who show a strong and irregular intraindividual variance in reading times, are corrected by using the estimates of the whole sample, and thus get a relatively low weight in the overall solution.

For the ML estimation, there are two different principles. *Full maximum likelihood* (FML) follows the target criterion to maximize the combined likelihood of regression coefficients (fixed parameters) and variance components. In *restricted maximum likelihood* (RML), the target criterion refers to the likelihood of the estimates for the variance components only (Raudenbush & Bryk, 2002), which means that the regression coefficients in the fixed part are not estimated. Thus, RML is combined with generalized least squares (GLS) estimation for the regression coefficients.

Which of the available techniques should be employed for estimating variance components? There are no clear guidelines under which circumstances FML or RML are to be preferred for estimating the variance components. FML may lead to biased (underestimated) variance components when the number of higher level units is small (Raudenbush & Bryk, 2002), but it has the advantage of allowing for hypothesis tests that compare the variance components of nested models differing in the fixed effects (see the Hypothesis Tests section). In most cases, FML and RML will produce highly similar results.

Which of the available techniques should be employed for estimating fixed effects? For the fixed effects, that is the regression coefficients in the fixed part of the model, simulation studies show that OLS, GLS, and ML techniques generally produce unbiased estimates, but that OLS seems to be less efficient than the other two methods, which means that far more observations are needed to produce precise estimates. For this reason, OLS cannot be recommended for small-scale research as reading time studies are likely to be. If intraclass correlation is present, OLS has the additional drawback that it leads to an underestimation of standard errors, thus yielding significance tests that are too liberal. In contrast, both the ML and GLS techniques generally result in accurate standard errors for fixed parameters (for overviews of simulation studies, see Hox, 1998; Kreft, 1996). Because reading time data tend to have a substantial intraclass correlation, ML or GLS are again to be preferred for the estimation of fixed effects.

Hypothesis Testing

Hierarchical linear models provide sophisticated options for hypothesis testing. Hypotheses concerning individual parameters can be tested by comparing the ratio of the parameter and its standard error to a standard normal distribution or to a student *t* distribution (Hox, 1998). Tests for random effects commonly use a statistic that follows a chi-square distribution (Raudenbush & Bryk, 2002, chap. 3). Akin to traditional linear models, hierarchical linear models also permit hypotheses concerning multiple parameters (in the form of a general linear hypothesis, Searle, 1971), which may be tested by a chi-square distributed statistic. Last but not least, there is the possibility of conducting individual and multiple parameter tests by comparing nested models. In nested models, one model is the "full" model, whereas another model is nested under the full model by fixing one or more parameters to zero. Significance tests are based on the chi-square distributed difference between the *deviances* of the two models. Deviance is an indicator of misfit, which is derived from the likelihoods of the ML estimates. Models differing in the fixed part can be compared only if FML estimates (but not RML estimates) were used for parameter estimation. Nested models have an important function in model building because the increment of more complex models can easily be tested against the fit of simpler models. By this means it is possible to test a sequence of nested models, starting from an unconditional model and stepwise proceeding from there on to ever more complex models that include sentence and person-level predictors.

Power and Required Sample Sizes

Similar to the situation in traditional linear models, the power of significance tests in hierarchical linear models is closely linked to the efficiency of parameter estimates: the smaller the variance of a particular estimate, the higher the probability that an invalid null hypothesis concerning this parameter will in fact be rejected. The power of significance tests in hierarchical linear models, however, depends on a complex pattern of factors: Apart from the magnitude of the effect in the population, the technique used for parameter estimation, the number of observations on both levels, and the intraclass correlation can make a difference (cf. Kreft & de Leeuw, 1998, chap. 5). Nevertheless, analytical considerations and the results of simulation studies allow at least some general suggestions as to how many observations will usually be needed to obtain sufficient power for simple two-level models.

For significance tests for level-1 coefficients, power increases with the total number of observations on level 1. Thus, the number of sentences is important for testing hypotheses concerning the influence of sentence-level predictors. Larger numbers of observations on level 1 are needed if the intraclass correlation is high. Likewise, for the power of significance tests for regression coefficients on level 2, large numbers of observations on this level are important. Thus, a reading time study investigating the influence of person variables should be based on a sufficiently high number of participants: It is not possible to compensate for a low number of observations on the person level by having each participant read more sentences. Based on simulation studies by Bassiri (1988) and van der Leeden and Busing (1994), Kreft (1996) proposed the “30/30 rule” for simple two-level models with cross-level interactions and ML estimates. According to the 30/30 rule, sufficient power for significance tests, including those for detecting cross-level interactions, is given if at least 30 observations on level 1, which are nested within 30 units on level 2, are available. There is, however, some evidence from simulation studies that estimates of fixed coefficients as well as variance components on level 2 are more efficient if the level-1 observations are distributed over many level-2 units, rather than having few level-2 units containing many level-1 observations (e.g., Mok, 1995). Moreover, efficient estimates of variance components on level 2 seem to require larger samples on level 2 than estimates of fixed effects (e.g., Maas & Hox, 2002). Based on these and similar findings, Hox (1998) has recommended two different rules of thumb for studies concerned with fixed effects including cross-level interactions and studies concerned with variance components. According to the “50/20” rule, cross-level interactions should be tested with at least 50 units on level 2, each with an average size of 20 observations on level 1. For studies concerned with variance components on level 2, Hox (1998) has proposed an even more asymmetric “100/10” rule, with 100 level-2 units and 10 level-1 observations nested within each level-2 unit.

In reading time research, the theoretical interest will often be set on the fixed coefficients rather than the variance components. The 50 participants called for by Hox's (1998) 50/20 rule correspond to typical sample sizes in experiments in text comprehension. In most studies, participants will read far more than 20 sentences, which leads to more efficient estimates for sentence-level coefficients. Hence, it is unlikely that the sample sizes required for conducting multilevel studies will impose extra time and costs on researchers. Some cautionary remarks are appropriate, though. First of all, the simulation studies carried out so far are based on rather simple models, with few parameters on each level. An especially high number of error components on the person level can dramatically increase the sample size needed because in addition to the variances, the covariances of the error components have to be estimated, too. Furthermore, as explained earlier, for various reasons such as a high intraclass correlation, a small effect size in the population, or multicollinearity of the predictor variables, a larger sample size may be needed to detect a particular effect. Precise rules for computing optimal sample sizes are not yet available for hierarchical linear models.

CONCLUSION

The first aim of this article was to demonstrate that hierarchical linear models are a viable statistical tool for the analysis of reading time data. In many situations, hierarchical linear models are superior to ANOVA or MR techniques. They take the multilevel structure of reading time data into account and avoid methodological problems associated with the traditional single-level methods. Although developed in the substantially different context of educational and social research, hierarchical linear models also open up new perspectives for psycholinguistic and text comprehension research by providing straightforward ways to test cross-level interactions of person and sentence characteristics. For this reason, multilevel models promise fruitful applications especially in research following a constructionist approach where assumptions concerning interactions of person characteristics (including reader expectations and reading goals) and text characteristics form the core of constructionist theories of text comprehension (e.g., Graesser, Singer, & Trabasso, 1994). The second aim of this article was to provide text comprehension researchers with some basic knowledge necessary to apply hierarchical linear models to reading time data. Despite the relative complexity of some aspects of the method (like model building or parameter estimation), it presents a workable—and often better—alternative to the statistical methods commonly used in reading time research. Sufficient power is attainable with sample sizes comparable to those of typical reading time studies.

There are a number of accessible introductory texts on hierarchical linear models, both concise primers (Kreft & de Leeuw, 1998; Nezlek, 2001, 2003) and comprehensive monographs (Goldstein, 2003; Raudenbush & Bryk, 2002; Snijders & Bosker, 1999). Blozis and Cudeck (1999) extended hierarchical linear models to handle latent variables at the second level. Because measurements of individual differences are often associated with large measurement error, the possibility of including latent variables at the person level is an attractive extension for researchers who are interested primarily in individual differences (e.g., reading skills) and their relationship to reading times. In addition to good introductory texts, user-friendly computer programs are available, providing many analysis options. The most common computer programs are HLM 6 (Raudenbush, Bryk, Cheong, & Congdon, 2004; see also http://www.ssicentral.com/hlm/) and MLwiN (Rasbash et al., 2000; see also http://www.mlwin.com/). A program called MPlus (Muthén & Muthén, 2003; see also http://www.statmodel.com/) allows multilevel modeling with latent variables on the second level.

From a theoretical perspective, it would even be valuable to enlarge the multilevel perspective by including intermediate levels between persons and sentences. All major theories of text comprehension presuppose that the processing of individual text segments depends partly on how the preceding text has been processed. Semantic and syntactic integration of words within sentences as well as many types of inferences (e.g., local and global bridging inferences) are examples for comprehension processes that include words or sentences as components of larger semantic structures. Accordingly, the allocation of processing resources to different sentences and, thus, their reading times may vary with the semantic context in which a sentence is encountered. As a consequence, sentence reading times collected for naturalistic texts are invariably structured by multiple levels beyond the distinction of sentence level and person level: Not only sentence characteristics but also properties of higher level semantic structures such as the paragraph or the text in which a sentence is encountered may contribute to variance in reading times. Reading times recorded for individual words or phrases imply an additional level of analysis that is nested within sentences. Provided that a data set contains sufficient information, hierarchical linear models permit investigations that include these additional levels of analysis.

ACKNOWLEDGMENTS

Completion of this article was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, Grant RI 1100/2–2). Thanks go to Inga Hoever and Sascha Schroeder for a critical reading of a previous version of the article.

REFERENCES

Aiken, S. A., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.

Barcikowski, R. S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6,* 267–285.

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model*. Unpublished PhD thesis, Department of Counseling, Educational Psychology and Special Education, Michigan State University.

Bisanz, G. L., Das, J. P., Varnhagen, C. K., & Henderson, H. R. (1992). Structural components of reading times and recall for sentences in narratives: Exploring changes with age and reading ability. *Journal of Educational Psychology, 84,* 103–114.

Blozis, S. A., & Cudeck, R. (1999). Conditionally linear mixed-effects models with latent variable covariates. *Journal of Educational & Behavioral Statistics, 24,* 245–270.

Burstein, L., Linn, R. L., & Capell, F. J. (1978). Analyzing multilevel data in the presence of heterogeneous within-class regressions. *Journal of Educational Statistics, 3,* 347–383.

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

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analyses for the behavioral sciences* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Cronbach, L. J., & Webb, N. (1975). Between class and within class effects in a reported aptitude X treatment interaction: A reanalysis of a study by G. L. Anderson. *Journal of Educational Psychology, 67,* 717–724.

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York: Oxford University Press.

Graesser, A. C., Hoffman, N. L., & Clark, L. F. (1980). Structural components of reading time. *Journal of Verbal Learning and Verbal Behavior, 19,* 135–151.

Graesser, A. C., Magliano, J. P., & Haberlandt, K. (1994). Psychological studies of naturalistic text. In H. van Oostendorp & R. A. Zwaan (Eds.), *Naturalistic text comprehension* (pp. 9–33). Norwood, NJ: Ablex.

Graesser, A. C., & Riha, J. R. (1984). An application of multiple regression techniques to sentence reading times. In D. E. Kieras & M. Just (Eds.), *New methods in reading comprehension research* (pp. 183–218). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review, 101,* 371–395.

Haberlandt, K. (1984). Components of sentence and word reading times. In D. E. Kieras & M. Just (Eds.), *New methods in reading comprehension research* (pp. 219–251). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Haberlandt, K. (1994). Methods in reading research. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 1–31). San Diego, CA: Academic Press.

Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114,* 357–374.

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

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis and data highways* (pp. 147–154). New York: Springer.

Juhasz, B. J., & Rayner, K. (2003). Investigating the effects of a set of intercorrelated variables on eye fixations during reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1312–1318.

Kreft, I. (1996). *Are multilevel techniques necessary? An overview, including simulation studies*. Retrieved April 23*,* 2004, from http://www.calstatela.edu/faculty/ikreft/quarterly/quarterly.html

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

Kreft, I., de Leeuw, J., & Aiken, L. S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research, 30,* 1–21.

Long, D. L., & Chong, J. L. (2001). Comprehension skill and global coherence: A paradoxical picture of poor comprehenders' abilities. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 27,* 1424–1429.

Longford, N. T. (1989). Fisher scoring algorithm for variance component analysis of data with multilevel structure. In R. D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 311–317). San Diego, CA: Academic Press.

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

Magliano, J. P., & Graesser, A. C. (1991). A three-pronged method for studying inference generation in literary text. *Poetics, 20,* 193–232.

Maas, C. J. M., & Hox, J. J. (2002). Sample sizes for multilevel modeling. In J. Blasius, J. Hox, E. de Leeuw, & P. Schmidt (Eds.), *Social science methodology in the new millennium: Proceedings of the Fifth International Conference on Logic and Methodology* (2nd, exp. ed.) [CD-ROM]. Opladen, Germany: Leske + Budrich.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology, 91,* 615–629.

Mok, M. (1995). Sample-size requirements for 2-level designs in educational research. *Multilevel Modelling Newsletter, 7*(2), 11–15.

Muthén, L., & Muthén, B. (2003). *Mplus user's guide*. Los Angeles: Muthén & Muthén.

Nezlek, J. B. (2001). Multilevel random coefficient analyses of event and interval contingent data in social and personality psychology research. *Personality and Social Psychology Bulletin, 27,* 771–785.

Nezlek, J. B. (2003). Using multilevel random coefficient modeling to analyze social interaction diary data. *Journal of Social and Personal Relationships, 20,* 437–469.

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

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., & Healy, M. (2000). *A user's guide to MLWiN* (2nd ed.). London: Institute of Education.

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin, 114,* 510–532.

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

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T. (2004). *HLM 6: Hierarchical linear and nonlinear modeling*. Chicago: Scientific Software International.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15,* 351–357.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

Singer, M. (1994). Discourse inference processes. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 479–515). San Diego, CA: Academic Press.

Snijders, T. A. B. (in press). Diagnostic checks for multilevel models. In J. de Leeuw & I. Kreft (Eds.), *Handbook of quantitative multilevel analysis*.

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

Stine-Morrow, E. A. L., Millinder, L., Pullara, O., & Herman, E. (2001). Patterns of resource allocation are reliable among younger and older readers. *Psychology & Aging, 16,* 69–84.

van der Leeden, R. (1998). Multilevel analysis of repeated measures data. *Quality and Quantity, 32,* 15–29.

van der Leeden, R., & Busing, F. M. T. A. (1994). *First iteration versus final IGLS/RIGLS estimates in two-level models: A Monte Carlo study with ML3* (Tech. Rep. No. PRM-02–94). Leiden, The Netherlands: Leiden University, Department of Psychometrics and Research Methodology.

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

Wisher, R. A. (1976). The effects of syntactic expectations during reading. *Journal of Educational Psychology, 68,* 597–602.

Wong, G. Y., & Mason, W. M. (1985). The hierarchical logistic regression model for multilevel analysis. *Journal of the American Statistical Association, 80,* 513–524.

Zwaan, R. A. (1994). Effect of genre expectations on text comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 920–933.