ELSEVIER

Available online at www.sciencedirect.com

SCIENCE DIRECT®

Speech Communication 43 (2004) 103–121

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

On multi-level modeling of data from repeated measures designs: a tutorial

Hugo Quené *, Huub van den Bergh

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

Received 28 February 2003; received in revised form 20 February 2004; accepted 23 February 2004

Abstract

Data from repeated measures experiments are usually analyzed with conventional ANOVA. Three well-known problems with ANOVA are the sphericity assumption, the design effect (sampling hierarchy), and the requirement for complete designs and data sets. This tutorial explains and demonstrates multi-level modeling (MLM) as an alternative analysis tool for repeated measures data. MLM allows us to estimate variance and covariance components explicitly. MLM does not require sphericity, it takes the sampling hierarchy into account, and it is capable of analyzing incomplete data. A fictitious data set is analyzed with MLM and ANOVA, and analysis results are compared. Moreover, existing data from a repeated measures design are re-analyzed with MLM, to demonstrate its advantages. Monte Carlo simulations suggest that MLM yields higher power than ANOVA, in particular under realistic circumstances. Although technically complex, MLM is recommended as a useful tool for analyzing repeated measures data from speech research.

AMS: 91-01; 62-01; 62-07

PACS: 43.10.Ln; 43.64.Yp; 43.66.Yw; 43.70.Jt

Keywords: Multi-level modeling; Repeated measures; Experimental design; Analysis of variance; Mixed effects; Variance components

1. Introduction

Like other behavioral disciplines, the study of speech communication progresses mainly by means of statistical inference. Using strict tools and procedures, researchers generalize from a sample of observed cases to broader contexts. Statistics as a discipline aims to facilitate and improve this inference, and to ensure validity of the resulting insights. Meanwhile, however, new insights are also achieved within the field of statistics itself, although these insights do not always percolate into actual research practice (Max and Onghena, 1999). Statistical insight and actual research practice thus are at risk to diverge, to the detriment of the latter. In particular, multi-level modeling (hence MLM) has emerged in the past decades as a highly flexible and useful tool for statistical analysis and inference (Searle et al., 1992; Bryk and Raudenbush, 1992; Goldstein, 1995; Hox, 1995; Kreft and De Leeuw, 1998; Snijders and Bosker, 1999; McCulloch and Searle, 2001; Raudenbush and Bryk, 2002; Maxwell and Delaney, 2004, Chapter 15). This new tool is

* Corresponding author. Tel.: +31-30-2536070; fax: +31-30-2536000.

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

URL: http://www.let.uu.nl/~Hugo.Quene/personal/.

0167-6393/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2004.02.004

also known as the hierarchical linear model, variance component model, or mixed-effects model. MLM has already found wide deployment in disciplines such as sociology (e.g. Carvajal et al., 2001), education (e.g. Broekkamp et al., 2002), biology (e.g. Agrawal et al., 2001; Hall and Bailey, 2001), and medicine (e.g. Beacon and Thompson, 1996; Merlo et al., 2001; Lochner et al., 2001). Its many advantages have also made MLM increasingly popular in behavioral research (e.g. Van der Leeden, 1998; Reise and Duan, 2001; Raudenbush and Bryk, 2002), but so far MLM has made few inroads into speech research.

The purpose of this tutorial is to explain the basics of multi-level modeling, to compare MLM against its more conventional counterpart for hypothesis testing, ANOVA, and to demonstrate the use of MLM in actual research in our field. Some readers might hesitate to learn about, let alone adopt statistical innovations such as MLM. It will be argued below that the advantages of MLM for inference and insight outweigh these difficulties.

The outline of this tutorial is as follows. First, three well-known problems with ANOVA are reviewed, using a fictitious data set. Multi-level modeling promises to solve all three problems: sphericity, hierarchical sampling, and missing data. The subsequent section explains the basics of multi-level modeling, using the same fictitious data set. Analysis results from MLM and from RM-ANOVA, based on the same data set, are then compared and discussed. One notable advantage of MLM is its higher power in hypothesis testing. MLM is then demonstrated in an example analysis of real data from a recently published study. MLM and ANOVA are also compared in a more general fashion, using Monte Carlo simulations. Finally, we discuss the advantages and drawbacks of using multi-level models for research in speech communication.

2. Three problems with ANOVA

2.1. Sphericity

This tutorial focuses on repeated-measurement designs, with two nested random factors: subjects (or participants), and trials (or occasions) within subjects, respectively. Hence, data are obtained from a *multi-level* sampling scheme, in which subjects have been sampled first, and trials have been sampled within subjects. These two levels of sampling are usually called level-1 (lower) and level-2 (higher). The factor of interest, like "treatment" or "condition", constitutes a fixed factor, so random and fixed effects are *mixed* in one experimental design.

To illustrate this design, we could think of a fictitious study of lip displacement during speaking, with $N = 108$ observations. There are $J = 12$ subjects in this study. Each subject participates in three trials or observations or replications in each of the three treatment conditions, yielding $n = 9$ observations for each subject. This design can be regarded as a two-level sample: first subjects are sampled from the population of suitable subjects, and then trials or observations are sampled from the possible observations within these sampled subjects. Stated otherwise, the observations are not all independent because observations are clustered within subjects; such observations within the same subject tend to be correlated. The fictitious measurements [1] for this study are given in Table 1, which shows the displacement data for 12 speakers (rows) under three treatment conditions (columns), with three trials or observations in each design cell.

The conventional analysis of these data would resort to repeated measures ANOVA (hence RM-ANOVA). The conventional univariate RM-ANOVA F test uses the Treatment by Subject interaction (with 2×11 df) as error term for the Treatment effect. This yields a significant main effect of Treatment: $F(2, 22) = 3.58$, $p = 0.045$. Hence we might conclude that treatments differ significantly in this study.

However, this would be an incorrect analysis for the present data set. A complication in all repeated measures designs is that observations

[1] These data were generated by adding a random effect (sampled from a normal distribution) to the fixed treatment effect. The data set is available online at the electronic appendix to this tutorial, at URL http://www.let.uu.nl/~Hugo.Quene/personal/multilevel.

Table 1
Fictitious data from a study of lip displacement, broken down by treatment conditions, by subjects (rows) and by trials (columns)

	Treatment A			Treatment B			Treatment C			
	1	2	3	4	5	6	7	8	9	$\overline{Y}_j$
1	1.71	−0.55	−0.75	0.54	1.07	0.01	1.32	2.94	2.41	0.97
2	−0.10	−0.68	1.07	0.48	1.72	1.73	2.43	1.23	1.82	1.08
3	−0.16	−0.75	−0.42	−2.40	−1.13	−4.74	1.67	0.86	1.65	−0.60
4	2.83	0.95	1.49	1.82	2.81	2.33	3.25	1.94	0.87	2.03
5	−1.15	−1.22	−0.88	1.48	0.75	−1.20	−1.32	0.58	1.28	−0.19
6	0.71	−0.68	1.58	−0.65	−0.32	1.55	1.72	3.52	2.52	1.10
7	−3.99	−2.69	−2.55	1.32	1.15	2.11	−3.61	−3.48	−2.32	−1.56
8	−3.25	−2.04	−4.50	−2.78	−1.77	−3.28	−3.03	−2.94	−2.13	−2.86
9	1.05	0.70	−0.13	0.58	−0.07	−1.52	0.77	2.62	1.53	0.61
10	0.19	−1.16	0.66	−2.54	−2.32	−3.07	2.42	2.57	1.80	−0.16
11	0.90	3.40	2.02	1.88	1.91	−0.03	4.77	5.76	3.73	2.70
12	1.02	0.78	0.12	−1.58	−1.20	−2.51	1.76	−0.18	−0.73	−0.28
$\overline{Y}$		−0.18			−0.22			1.11		

Measurements are given in arbitrary units.

within each subject are correlated, precisely because they are made within the same subject. There is of course variance between multiple trials or occasions for the same treatment, but there is also covariance between occasions under different treatments. One can think of this covariance as an unstandardized amount of *correlation* between the scores in two treatments. The variance–covariance matrix in (1) lists the variances (on the diagonal) and the covariances (off the diagonal) between treatments A, B, and C.

$$\begin{bmatrix} \sigma^2_{AA} & \mathrm{cov}_{AB} & \mathrm{cov}_{AC} \\ \mathrm{cov}_{BA} & \sigma^2_{BB} & \mathrm{cov}_{BC} \\ \mathrm{cov}_{CA} & \mathrm{cov}_{CB} & \sigma^2_{CC} \end{bmatrix} \tag{1}$$

If all variances on the diagonal have the same value, then there is homogeneity of variance, which is a necessary condition to perform RM-ANOVA. In addition to this property, conventional RM-ANOVA also requires that all covariances off the diagonal have the same value; this property is called *compound symmetry*. Compound symmetry is a sufficient condition to perform univariate RM-ANOVA (Winer, 1971, p. 596). If the stringent assumption of compound symmetry is violated, then the weaker assumption of *sphericity* must still hold, as a necessary condition to perform univariate RM-ANOVA (Maxwell and Delaney, 2004, Chapter 11). Sphericity is found when all difference scores between pairs of treatments have the same variance. [2] In other words, the within-subject correlation of scores is properly accounted for in RM-ANOVA, but only if the assumption holds that these correlations (covariances) are equal among all treatments conditions. After factoring out the subjects' individual differences, residual scores must have homogeneous variances among all cells of the subject by treatment interaction. This assumption, however, is often violated (O'Brien and Kaiser, 1985; Max and Onghena, 1999).

For the present data set, there is indeed no sphericity (Mauchly's $W = 0.317$, $p = 0.003$), and hence a univariate RM-ANOVA is not appropriate. If the sphericity assumption is violated, then there are several possibilities. The first is to reduce the degrees of freedom in the univariate F test, which results in more conservative testing than with uncorrected degrees of freedom as above. In this example, we use the Huynh–Feldt correction factor $\epsilon = 0.625$. The resulting univariate RM-ANOVA F test with corrected degrees of freedom does *not* yield a significant main effect of

[2] The variance of a difference score between treatments A and B is defined as $\mathrm{var}(A - B) = \mathrm{var}(A) + \mathrm{var}(B) - 2\ \mathrm{cov}(A, B)$. Hence, the sphericity assumption implicitly refers to the covariances between pairs of treatments.

Treatment: $F(1.25, 13.75) = 3.58$, $p = 0.073$. The previous value of $F(2, 22) = 3.58$ with uncorrected degrees of freedom was positively biased towards finding significant effects, due to the violation of the sphericity assumption (O'Brien and Kaiser, 1985, and references given there).

The second possibility is to use a multivariate approach to RM-ANOVA (O'Brien and Kaiser, 1985; Max and Onghena, 1999; Maxwell and Delaney, 2004); this approach does not depend on the sphericity assumption. In our example, the multivariate RM-ANOVA F test yields a significant main effect of Treatment: $F(2, 10) = 10.4$, $p = 0.004$. This shows that the multivariate F test has a greater power than the univariate F test, if the sphericity assumption is violated. The Monte Carlo simulations below further address this power issue.

Max and Onghena (1999) also mention MLM as a third option if the sphericity assumption is violated. In solving the sphericity problem, MLM has the important advantage that it allows the modeling of the variance–covariance matrix (1) directly from the observed data. No assumptions are necessary about constant variances (homogeneity of variance, or homoschedasticity) nor about constant covariances (compound symmetry) nor about constant variances of difference scores (sphericity). This means that MLM can be used safely, even if there is no homogeneity of variance, or if there is no sphericity, or if there is no compound symmetry (Goldstein, 1995; Beacon and Thompson, 1996). Moreover, the three ANOVA assumptions mentioned above impose hard limits on the insights that can be extracted from a data set. Rather than *assuming* that variances are equal and that covariances are equal, it would be more advantageous to *inspect* estimates of these quantities, under various treatment conditions. MLM allows researchers to do this, as will be shown below with real data. The variances and covariances under different treatments often tell their own tale about the research question at hand. In spite of these advantages, however, Max and Onghena (1999) mention MLM only briefly, and without further explanation. The present tutorial provides more explanation, background and examples of multi-level modeling.

2.2. Design effect

In a multi-level sampling procedure, such as in most repeated measures experiments, the estimated standard error is usually larger than in a single-level sampling procedure with the same number of elements (Kish, 1967; Cochran, 1977). Due to correlation (covariance) of observations within higher-level units, the effective number of observations is lower than the total number of observations.

In the example above, the observations within a subject are correlated. This within-subject correlation is generally known as *intra-class correlation*; it is also related to the non-centrality parameter λ. In this example, the between-subjects variance is 2.056; the residual within-subject variance is 2.408. The intra-class correlation ρ_I is then computed as $2.056/(2.056 + 2.408) = 0.461$ (Haggard, 1958; Winer, 1971; Snijders and Bosker, 1999, p. 20). This moderate intra-class correlation coefficient indicates that observations are indeed clustered or correlated within subjects.

The effective number of observations, therefore, is less than the nominal $N = 108$. For this two-stage sampling design, the so-called *design effect* is defined as $1 + (n - 1)\rho_I \approx 4.7$ (Kish, 1967; Cochran, 1977; Snijders and Bosker, 1999). The effective sample size in this study is not $N = 108$, but N divided by the design effect, or $108/4.7 \approx 23$. Hence, the effective standard error of the mean (defined as $s/\sqrt{N}$) is not $2.076/\sqrt{108} = 0.200$, but $2.076/\sqrt{23} = 0.433$, more than twice as large.

Data sets with a multi-level structure can of course be analyzed as if they were drawn from a single-level sample at the lower level, disregarding the higher-level sampling structure. This so-called disaggregation is sometimes found in speech research using repeated measures [3] (Snijders and Bosker, 1999; Max and Onghena, 1999). Such disaggregation of higher-level units into a single-level model leads to underestimation of the standard error of the mean, as explained above. In our example the unbiased estimate for the standard

[3] In the present example, disaggregated ANOVA would yield $F(2, 105) = 5.15$, $p = 0.007$ for the treatment effect.

error of the mean is 0.433; the disaggregated estimate is 0.200, which is far too low. Most statistical analyses use (quantities derived from) this estimated standard error of the mean. Hence, disaggregation results in an underestimation of the Type I error: H_0 would be rejected too easily while in fact H_0 may be true (Snijders and Bosker, 1999).

Speech researchers generally take the intra-class correlation into account, however, and do not disaggregate their data. RM-ANOVA is widely used in speech research; this technique allows researchers to separate random between-subject variance from random within-subject variance. Only the latter is then used to test within-subject factors. This decreases chance capitalization, which would be a serious risk if data were disaggregated.

Although RM-ANOVA is very useful for such a two-level sampling procedure, it cannot be extended to higher-order sampling procedures. To illustrate this point, we could further imagine that our fictitious lip displacement study is conducted in 15 languages, which are randomly selected from the thousands of languages of the world. As before, 12 speakers participate for each language. (A plausible aim of this research project could be to investigate whether the treatment effect occurs universally across languages.) RM-ANOVA cannot be used to analyze data from such a three-level sampling procedure involving languages, speakers within languages, and trials within speakers within languages. Because of the lack of suitable statistical techniques, researchers have been forced until now to avoid such quantitative investigations.

2.3. Missing data

A third considerable problem in RM-ANOVA is that missing data are not allowed. Disentangling the between-subject and between-treatment variances requires a full data matrix. The first consequence of this requirement is that if a subject misses one observation (perhaps for good reasons), then all data from that subject have to be discarded. This threatens the power of the experimental design. A common strategy is to replace missing observations with a subject's mean, but this strategy results in underestimation of the error variance.

Second, the ban on missing observations hampers the analysis of incomplete designs, such as Latin-squares. Many experiments require counterbalancing of subjects across treatments, to neutralize strategic effects or learning effects. RM-ANOVA of such designs is notoriously complex, and usually requires an auxiliary factor such as List or Group in the analysis (cf. Raaijmakers et al., 1999). The variances attributed to this auxiliary factor and its interactions should in fact be regarded as between-subject variance. Multi-level modeling provides an attractive alternative analysis, because it allows statistical evaluation of incomplete data, without any additional complication (Snijders and Bosker, 1999, p. 170).

3. Multi-level modeling

Multi-level modeling promises to solve all three problems with conventional RM-ANOVA discussed above. It is robust against violations of homoschedasticity and sphericity. It is suitable for analyzing data from multi-level sampling schemes. It is also robust against missing data. But how does it work? This section explains the basics of multi-level modeling. Several comprehensive textbooks are available for further study, including Goldstein (1995), Bryk and Raudenbush (1992), Hox (1995), Kreft and De Leeuw (1998), Snijders and Bosker (1999) and Raudenbush and Bryk (2002).

3.1. Constructing a model

Suppose that Y_{ij} is a response variable, for trial i within subject j. This response can be regarded as a deviation from the mean B_j of the jth subject, i.e.

$$Y_{ij} = B_j + e_{ij} \tag{2}$$

It is assumed that the residuals e_{ij} are normally distributed, with mean zero and variance $\sigma^2_{e_{ij}}$. We can also regard the mean response B_j for subject j as a deviation from the grand mean γ_{00}, i.e.

$$B_j = \gamma_{00} + u_{0j} \tag{3}$$

Just like residuals e_{ij} at the lowest level (of trials), the residuals u_{0j} at the higher level (of subjects) are assumed to be normally distributed, with an expected value of zero and with variance $\sigma^2_{u_{0j}}$. Higher-level residuals u_{0j} are also assumed to be uncorrelated with lower-level residuals e_{ij}, i.e. $r_{u_{0j},e_{ij}} = 0$. Substitution of Eq. (3) in (2) yields the basic multi-level model:

$$Y_{ij} = \gamma_{00} + (u_{0j} + e_{ij}) \tag{4}$$

where i ranges over $1,\dots,I_j$ level-1 units (e.g. trials or occasions), and j ranges over $1,\dots,J$ level-2 units (e.g. subjects).

This model consists of two parts, viz. the fixed part and the random part (in this paper, random terms in models are in parentheses). The fixed part here only models the "effect" of the grand mean that underlies all observations. In the random part, the total variance of Y is decomposed to two levels, viz. variance between subjects $\sigma^2_{u_{0j}}$, and between occasions within subject $\sigma^2_{e_{ij}}$. Because there are no explanatory variables, (4) is also called the "empty" model (Snijders and Bosker, 1999, p. 46). The corresponding variance–covariance matrix is given in (5).

$$\begin{bmatrix} \sigma^2_{u_{0j}} + \sigma^2_{e_{ij}} & 0 & 0 \\ 0 & \sigma^2_{u_{0j}} + \sigma^2_{e_{ij}} & 0 \\ 0 & 0 & \sigma^2_{u_{0j}} + \sigma^2_{e_{ij}} \end{bmatrix} \tag{5}$$

Variances on the diagonal of this matrix are constrained to be the same for all treatments, and all variances off the diagonal are set to zero (yielding compound symmetry). This matrix can therefore be reduced to:

$$\begin{bmatrix} \sigma^2_{u_{0j}} + \sigma^2_{e_{ij}} \end{bmatrix} \tag{6}$$

The empty model (4) can easily be extended by including explanatory variables. We begin by including a single explanatory variable in the fixed part. In the fictitious study of lip displacement, we may extend the model in (4) by including the Treatment factor as an explanatory variable. This can be done by replacing the pooled mean γ_{00} in (4) by three coefficients γ_{A00}, γ_{B00} and γ_{C00}, for the three treatment conditions. The selection of these conditions is done by means of three binary dummy variables, here named TrA, TrB, and TrC (e.g. Pedhazur and Schmelkin Pedhazur, 1991). The dummy variable TrA has values unity (1 = "on") for observations in treatment condition A and zero (0 = "off") for the other conditions; the other two dummy variables are constructed in analogous fashion. The resulting model in (7) below is also called the *cell means* model (Searle, 1987).

$$Y_{ij} = \gamma_{A00}\mathrm{TrA} + \gamma_{B00}\mathrm{TrB} + \gamma_{C00}\mathrm{TrC} + (u_{0j} + e_{ij}) \tag{7}$$

This model includes the treatment effects only in its fixed part. Since we have only adjusted the fixed part of the model, the variance–covariance matrices for the random components u_{0j} and e_{ij} are the same as specified in (5) above. The variance between subjects and within subjects is constrained to be the same for all three treatment conditions. This implies that the main effects of treatments are equal among all subjects and among all trials within subjects.

Since these assumptions may be invalid, the model could be further improved to allow for non-constant variances among the treatments, both between subjects and between trials within subjects, as in (8) below.

$$\begin{aligned} Y_{ij} &= \gamma_{A00}\mathrm{TrA} + \gamma_{B00}\mathrm{TrB} + \gamma_{C00}\mathrm{TrC} \\ &\quad + (u_{A0j}\mathrm{TrA} + u_{B0j}\mathrm{TrB} + u_{C0j}\mathrm{TrC} \\ &\quad + e_{Aij}\mathrm{TrA} + e_{Bij}\mathrm{TrB} + e_{Cij}\mathrm{TrC}) \\ &= \mathrm{TrA}[\gamma_{A00} + (u_{A0j} + e_{Aij})] \\ &\quad + \mathrm{TrB}[\gamma_{B00} + (u_{B0j} + e_{Bij})] \\ &\quad + \mathrm{TrC}[\gamma_{C00} + (u_{C0j} + e_{Cij})] \end{aligned} \tag{8}$$

Just like the grand mean γ_{00} has been replaced in the cell means model (7) by three treatment means, so the general variances both between subjects and within subjects are replaced here by three treatment variances, at each level of the sampling hierarchy. The variance–covariance matrix for the random variance at level-2 (between subjects) according to model (8) is given in (9). Variances on the diagonal do not need to be the same for each treatment.

$$\begin{bmatrix} \sigma^2_{u_{A0j}} & 0 & 0 \\ 0 & \sigma^2_{u_{B0j}} & 0 \\ 0 & 0 & \sigma^2_{u_{C0j}} \end{bmatrix} \tag{9}$$

This model no longer assumes homoschedasticity, but it still assumes compound symmetry or sphericity. As mentioned above, MLM does not require the sphericity assumption for valid inferences. Rather than constrain the covariances, MLM allows us to specify the full variance–covariance matrix (1) in the model. Due to the particular design of this study, however, we can only do so here for the variance component u_{0j} between subjects (level-2). There is no covariance between trials within subjects (level-1) in this example design, because trials were not co-varied with treatment conditions here. This yields the following fully specified multi-level model (10) with its corresponding variance–covariance matrix given in Eq. (11):

$$\begin{aligned} Y_{ij} &= \gamma_{A00}\mathrm{TrA} + \gamma_{B00}\mathrm{TrB} + \gamma_{C00}\mathrm{TrC} \\ &\quad + (u_{A0j}\mathrm{TrA} + u_{B0j}\mathrm{TrB} + u_{C0j}\mathrm{TrC} \\ &\quad + e_{Aij}\mathrm{TrA} + e_{Bij}\mathrm{TrB} + e_{Cij}\mathrm{TrC}) \\ &= \mathrm{TrA}[\gamma_{A00} + (u_{A0j} + e_{Aij})] \\ &\quad + \mathrm{TrB}[\gamma_{B00} + (u_{B0j} + e_{Bij})] \\ &\quad + \mathrm{TrC}[\gamma_{C00} + (u_{C0j} + e_{Cij})] \end{aligned} \tag{10}$$

$$\begin{bmatrix} \sigma^2_{u_{AA0j}} & \mathrm{cov}(u_{B0j}, u_{A0j}) & \mathrm{cov}(u_{C0j}, u_{A0j}) \\ \mathrm{cov}(u_{A0j}, u_{B0j}) & \sigma^2_{u_{BB0j}} & \mathrm{cov}(u_{C0j}, u_{B0j}) \\ \mathrm{cov}(u_{A0j}, u_{C0j}) & \mathrm{cov}(u_{B0j}, u_{C0j}) & \sigma^2_{u_{CC0j}} \end{bmatrix} \tag{11}$$

This most extensive model captures all relevant structure of the data set, both in its fixed and in its random part. Estimated coefficients for this and preceding models are presented and discussed below.

3.2. Estimating coefficients and testing hypotheses

The coefficients in multi-level models are calculated by means of a variety of computational estimation procedures, such as Iterated Generalized Least Squares (or IGLS, Goldstein, 1995), Fisher scoring (Longford, 1993) and Expectation-Maximization (or EM, Bryk and Raudenbush, 1992; Raudenbush and Bryk, 2002). All these estimation procedures are based on advanced matrix algebra.

Although multi-level models require fewer assumptions than classical ANOVA models, and less stringent ones, a few key assumptions are still necessary for the estimation procedures used in MLM. First, the random components (such as u_{0j} and e_{ij} in the models above) should have a normal distribution, with zero mean. In contrast to RM-ANOVA, no assumptions are made about the *variances* of these components. Hence MLM can be regarded as a generalization of ANOVA, without the problematic constraints discussed above. MLM does not require homoschedasticity, nor compound symmetry, nor sphericity. Second, the dependent variable should follow a normal distribution. Special techniques are available in case the dependent variable is not normally distributed, as with binary or discrete variables (Goldstein, 1991), but we will not discuss those complications here (see e.g. Snijders and Bosker, 1999, Chapter 14). These two assumptions are more realistic than the more stringent assumptions required by RM-ANOVA models. The procedures differ somewhat in their assumptions during estimation. IGLS and Fisher scoring assume that the residuals at all levels are randomly distributed; this assumption is required to calculate the variance–covariance matrices at each level. The EM procedure draws heavily on the joint distribution of both the observed response variables and the unknown residuals. Under the above two key assumptions (randomly distributed residuals, and normally distributed responses), however, all estimation procedures yield the same estimates for the coefficients in the multi-level model (Goldstein et al., 1994; Snijders and Bosker, 1999, Section 4.6). All estimation methods use iterative procedures. Each iteration consists of two steps. Fixed parameters are estimated first; then the random parameters are estimated in a second step. In the next iteration, the fixed parameters are estimated again, using the previously estimated variance–covariance matrices; then the random parameters are estimated again, using the previously estimated fixed parameters. These iterative estimation methods converge rapidly to the maximum likelihood estimates.

Several textbooks on MLM contain useful expositions of the computational procedures for estimating model coefficients (Goldstein, 1995; Bryk and Raudenbush, 1992; Hox, 1995; Kreft and De Leeuw, 1998; Snijders and Bosker, 1999; Raudenbush and Bryk, 2002). Pinheiro and Bates (2000, Chapter 2) and Raudenbush and Bryk (2002, Chapters 3 and 14), provide a helpful introduction and overview of the various estimation procedures. In the remainder of this tutorial, we will abstract from the computational details of the estimation procedures.

The resulting estimates can be used for hypothesis testing, e.g. to evaluate H_0 claiming that a particular estimated parameter equals zero. This type of hypothesis testing is based on the Wald criterion that an estimate is significant at $\alpha = 0.05$, if it exceeds 1.96 times its associated standard error (e.g. Hox, 1995). In other words, if the ratio of an estimated parameter and its standard error exceeds 1.96, then the H_0 regarding that parameter is rejected. The parameter then differs significantly from zero, and should be included in an adequate model of the data set.

Contrasts between coefficients are tested in a similar fashion. For example, the main effect of the treatment factor can be evaluated in the form of pairwise comparisons between treatment conditions. Just as in ANOVA comparisons, each contrast corresponds with an estimated amount of variance (with its associated standard error of that estimate). The variance quantity itself follows a χ^2 distribution (Winer, 1971, p. 849). Hence, a contrast is evaluated using χ^2 with $\mathrm{df} = 1$ as test statistic (for details see Raudenbush and Bryk, 2002; Goldstein, 1995, p. 33). If multiple pairwise comparisons are required (for more than two treatment conditions), as in our example study, then there are multiple contrasts to evaluate. The significance level for each separate contrast should then be adjusted, e.g. using Bonferroni adjustment of α divided by the number of comparisons (here $\alpha = 0.05/3 = 0.016$) to ensure the overall significance level across all pairwise comparisons (Kirk, 1995).

3.3. Results of the example model

The estimated parameters [4] from the models (4), (7) and (10) are given in Table 2 below.

Model (4) is the "empty" model, with the grand mean (0.237) as its only explanatory variable. Note that the estimated standard error of the mean (0.440) approximates the effective standard error corrected for the design effect (calculated above as 0.433). This estimate is indeed considerably larger than the single-level estimate (0.200). As we already saw above, both random variance components differ from zero (based on Wald testing: estimate divided by standard error exceeds 1.96). This shows that a multi-level model is indeed necessary. Lip displacement varies both between and within subjects, and these two sources of random variance should be taken into account. The estimate for the variance among trials (level-1) within subjects is more accurate (with smaller standard error of estimate), because this estimate is based on a large number of trials within subjects. The estimate for the variance among subjects (level-2), however, is based on only 12 subjects' means.

The empty model (4) is equivalent with a single-level one-way ANOVA with subjects (or generally, level-2 units) as the main effect. For the present data set, this yields $F(11, 96) = 8.69$, $p < 0.001$. The within-subject variance $s^2_{\mathrm{w}} = 2.408$ is an unbiased estimate of the population within-subject variance. An unbiased estimate of the population between-subject variance can be calculated as $(s^2_{\mathrm{w}}/n) \times (F - 1) = 2.058$ with $n = 9$ replications within each subject (Snijders and Bosker, 1999, p. 22). The estimated coefficients in Table 2 for the empty model are indeed very close to these unbiased variance estimates.

In model (7), the mean score is estimated for each treatment condition separately. Obviously, the first question is whether the differences between treatments are significant. As explained in the previous section, such hypotheses are tested by evaluating the variance attributed to the relevant

[4] Analyses were done with the `MLwiN` program (Rasbash et al., 2000). An annotated log of these example analyses is available at the electronic appendix to this tutorial, at URL http://www.let.uu.nl/~Hugo.Quene/personal/multilevel.

Table 2
Estimated parameters (with standard error of estimate in parentheses) of multi-level modeling of fictitious data from Table 1

	Model (4)	Model (7)	Model (10)
Fixed			
γ_{00}	0.237 (0.440)		
γ_{A00}		−0.180 (0.481)	−0.180 (0.474)
γ_{B00}		−0.219 (0.481)	−0.219 (0.518)
γ_{C00}		1.110 (0.481)	1.110 (0.638)
Random			
$\sigma^2_{u_{0j}}$	2.056 (0.944)	2.099 (0.944)	
$\sigma^2_{e_{ij}}$	2.408 (0.348)	2.020 (0.292)	
$\sigma^2_{u_{A0j}}$			2.411 (1.097)
$\sigma^2_{u_{B0j}}$			2.903 (1.312)
$\sigma^2_{u_{C0j}}$			4.617 (1.986)
$\mathrm{cov}(u_{A0j}, u_{B0j})$			0.926 (0.887)
$\mathrm{cov}(u_{A0j}, u_{C0j})$			3.300 (1.408)
$\mathrm{cov}(u_{B0j}, u_{C0j})$			0.826 (1.164)
$\sigma^2_{e_{Aij}}$			0.851 (0.246)
$\sigma^2_{e_{Bij}}$			0.964 (0.278)
$\sigma^2_{e_{Cij}}$			0.819 (0.236)
Deviance	426.3	407.5	355.8

contrast(s), with Bonferroni adjustment for the number of comparisons. Not surprisingly, given the treatment means in Table 1, the pairwise comparison between treatments *A* and *B* is not significant (*A*–*B*: $\chi^2 = 0.01$). The other two pairwise comparisons however are highly significant (*A*–*C*: $\chi^2 = 14.8$; *B*–*C*: $\chi^2 = 15.8$; both $p < 0.001$). Treatment *C* yields significantly higher scores than the other treatments.

The cell means model (7) corresponds with a univariate RM-ANOVA with Treatment as a factor. When the data were analyzed with RM-ANOVA with correction for violation of sphericity (see above), the main effect of Treatment was not significant. For these example data, MLM turns out to be more powerful, as it yields significant contrasts for the Treatment factor. The inverse fit or deviance (defined as $-2\log$ likelihood, Snijders and Bosker, 1999, Section 6.2) of each model is given in the bottom part of Table 2. The considerable reduction in deviance indicates that model (7) is indeed an improvement [5] over the empty model (4). This is what one would expect, because the within-subject variance can now partly be attributed to the fixed effect of Treatment.

The third model reviewed here does not make any assumptions about variances or covariances in the model. All these quantities are estimated, see Table 2. Hence, these quantities become properties to be modeled and understood, rather than a priori postulated (Goldstein et al., 1994). For example, we see that the variance between subjects is larger in treatment *C* than in the other two treatments—although the pairwise comparisons among between-subject variance components do not show significant contrasts (*A*–*B*: $\chi^2 = 0.09$, n.s.; *A*–*C*: $\chi^2 = 3.11$, $p = 0.078$; *B*–*C*: $\chi^2 = 0.54$, n.s.). Nevertheless, it would be inappropriate to assume homoschedasticity here. Subjects appear to be more different under treatment *C* than under other treatments.

The covariances also show interesting properties of the data set. There is high covariance at level-2 between treatments *A* and *C*: $\mathrm{cov}(u_{A0j}, u_{C0j}) = 3.300$ (with standard error of estimate 1.408). This covariance can be regarded as unstandardized correlation between subjects' means under treatments *A* and *C*. The standardized correlation coefficient, $r = (3.300)^2/(2.411 \times 4.617) = 0.98$, is extremely high. Subjects' means in treatments *A*

[5] The reduction in deviance between subsequent models can be evaluated using χ^2, but only if the two models compared have the same fixed parts, and differ only in their random parts (Snijders and Bosker, 1999; Pinheiro and Bates, 2000). Here we will only compare deviances informally.

and C are obviously highly correlated. Thus, we see that there is neither homoschedasticity nor sphericity nor compound symmetry in this data set. If these data had been obtained in a real experiment, then such differences in variances and covariances would have called for an explanation, either in theoretical terms or as a strategic effect induced by the experiment. For example, we know that conditions with higher average scores tend to have larger variances as well (Max and Onghena, 1999); this might explain the higher between-subject variance under treatment C.

In general, changing the random (co)variance components (as in model (7)) also yields changes in the estimated fixed effects or treatment means; this particular case is an exception in that the estimated treatment means do not change. However, the additional random components do affect the standard errors of these estimated treatment means, which in turn affects the test statistic used for evaluating the pairwise comparisons of the three treatments. For example, the test statistic for the pairwise comparison between treatment conditions B and C in the fixed part was $\chi^2 = 15.8$ in the previous cell means model (7) (with $\mathrm{df} = 1$, $p < 0.001$). In the present full model (10), however, the test statistic for this comparison has decreased considerably to $\chi^2 = 3.28$ (with $\mathrm{df} = 1, p = 0.070$). The difference between treatments B and C has disappeared, in terms of its significance, after we have taken into account the absence of homoschedasticity, and the presence of intercorrelations within subjects among their treatment means.

As argued above, this conservative behavior of MLM is entirely appropriate. In model (10), differences in level-2 variances cannot contribute to the comparison of conditions. In the predecessor model, as in the ANOVA model, this absence of homoschedasticity inflates the fixed contrasts and effects of interest. H_0 is then incorrectly rejected, although the true unbiased difference between conditions B and C is in fact not significant.

4. Multi-level modeling of existing data

As a further demonstration, let us apply the multi-level modeling technique outlined above to real data, to show what could be gained by using such modeling in actual research. To this end, we have re-analyzed data from a recent prosody study with 9 esophageal, 10 tracheoesophageal and 10 laryngeal control speakers (Van Rossum et al., 2002, Experiment 2). The 29 speakers read 10 sentences. Each speaker read each sentence twice, with two different preceding sentences that induced a contrastive accent either on a critical word early in the sentence, or on a different word later in the sentence. Hence, the critical word in each sentence was produced both with and without contrastive accent. For our present purposes, the acoustic duration of this critical word (expressed in ms units) constitutes the dependent variable.

The main hypothesis in this re-analysis is that esophageal and tracheoesophageal speakers, who cannot use F_0 to signal accent, make larger durational contrasts to signal accent than laryngeal control speakers do. The absence of F_0 movements, which are the most important cues to signal accent (Sluijter, 1995), is compensated for by enhanced durational cues. Van Rossum et al. (2002) reject this hypothesis, however, on the basis of a qualitative analysis of the acoustic measurements, after the acoustic measurements are reduced to nominal factors (“speaker uses duration cue”—“speaker does not use duration cue”).

In a conventional Repeated Measures ANOVA, the above hypothesis could have been tested by means of an interaction effect between speaker group (between speakers) and accent condition (within speakers). This interaction among the 3×2 cells of this design is indeed significant in a repeated measures ANOVA, $F(2, 26) = 7.19$, $p = 0.003$, supporting the main hypothesis (both main effects were also highly significant: speaker group, $F(1, 26) = 16.36$, $p < 0.001$; accent, $F(1, 26) = 101.35$, $p < 0.001$).

As explained above, however, such an ANOVA would have been based on several questionable assumptions. In particular, the research hypothesis predicts that the assumption of compound symmetry is violated, because it states that speaker groups differ in their ability to use duration to signal accent. It is entirely likely that speaker groups differ not only in their *average* word durations in the accent conditions, but also in their *variances* in word durations in the accent conditions. Hence, compound symmetry is likely to be absent, and sphericity is threatened a priori. Fig. 1 summarizes the speakers' average word durations for accented and unaccented words. This illustrates that durations among the two accent conditions are highly correlated for tracheoesophageal and for laryngeal control speakers, but not for esophageal speakers. The assumption of compound symmetry (equal covariances) among speaker groups is therefore not warranted, nor is the assumption of homoschedasticity.

4.1. Re-analysis of discrete factors

For expository reasons we start our re-analysis with the cell means model, taking the multi-level sampling hierarchy into account. This is done by estimating six coefficients (γ_{EU00}, γ_{EA00}, etc.) for the six main cells of this design. The appropriate selection is done by means of six binary dummy variables, here named EU (Esophageal Unaccented), EA (Esophageal Accented), TU and TA (Tracheoesophageal) and NU and NA (Normal control speakers), similar to model (7) above. This yields the cell means model in (12):

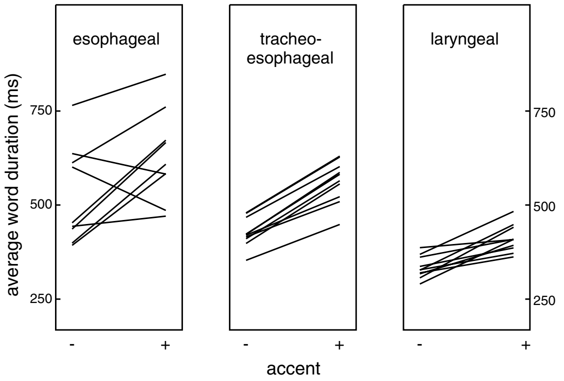

Fig. 1. Speakers' average word durations in ms, broken down by speaker group and accent condition. Data are from the study reported by Van Rossum et al. (2002).

$$\begin{aligned}D_{ij} = {} & \text{EU}_{ij}\gamma_{\text{EU00}} + \text{EA}_{ij}\gamma_{\text{EA00}} + \text{TU}_{ij}\gamma_{\text{TU00}} \\ & + \text{TA}_{ij}\gamma_{\text{TA00}} + \text{NU}_{ij}\gamma_{\text{NU00}} + \text{NA}_{ij}\gamma_{\text{NA00}} \\ & + (u_{0j} + e_{ij}) \end{aligned} \qquad (12)$$

Coefficients for this cell means model (12) will not be estimated here, however, because the model is not appropriate, for two reasons. First, variances between speakers ($\sigma^2_{u_{0j}}$) and within speakers ($\sigma^2_{e_{ij}}$) are assumed to be equal for the six main cells (homoschedasticity at both levels). Second, the correlations between the two accent conditions are assumed to be equal for the three speaker groups (compound symmetry at level-2). As explained above, both properties of the data set should be investigated in their own right, rather than postulated a priori. Model (13) achieves this, by including separate random estimates for each sampling level for each main cell. In addition, between-speaker covariance terms are included at level-2 (speakers) only [6] as shown in the corresponding variance–covariance matrix in (14).

$$\begin{aligned}D_{ij} = {} & \text{EU}_{ij}[\gamma_{\text{EU00}} + (u_{\text{EU0}j} + e_{\text{EU}ij})] \\ & + \text{EA}_{ij}[\gamma_{\text{EA00}} + (u_{\text{EA0}j} + e_{\text{EA}ij})] \\ & + \text{TU}_{ij}[\gamma_{\text{TU00}} + (u_{\text{TU0}j} + e_{\text{TU}ij})] \\ & + \text{TA}_{ij}[\gamma_{\text{TA00}} + (u_{\text{TA0}j} + e_{\text{TA}ij})] \\ & + \text{NU}_{ij}[\gamma_{\text{NU00}} + (u_{\text{NU0}j} + e_{\text{NU}ij})] \\ & + \text{NA}_{ij}[\gamma_{\text{NA00}} + (u_{\text{NA0}j} + e_{\text{NA}ij})] \end{aligned} \qquad (13)$$

Although this model might look somewhat intimidating, its interpretation is rather straightforward. The six fixed parameters γ_{EU00} to γ_{NA00} correspond to the respective cell means. The random components u_{0j} and e_{ij} are also estimated

[6] Again, covariances are only specified between accent conditions at level-2 (subjects). The grouping factor is outer to the speaker factor (or, speakers are nested within groups), and hence there is no covariance between groups at the speaker level.

$$\begin{bmatrix} \sigma^2_{u_{\mathrm{EU}0j}} & & & & & \\ \mathrm{cov}(u_{\mathrm{EU}0j}, u_{\mathrm{EA}0j}) & \sigma^2_{u_{\mathrm{EA}0j}} & & & & \\ 0 & 0 & \sigma^2_{u_{\mathrm{TU}0j}} & & & \\ 0 & 0 & \mathrm{cov}(u_{\mathrm{TU}0j}, u_{\mathrm{TA}0j}) & \sigma^2_{u_{\mathrm{TA}0j}} & & \\ 0 & 0 & 0 & 0 & \sigma^2_{u_{\mathrm{NU}0j}} & \\ 0 & 0 & 0 & 0 & \mathrm{cov}(u_{\mathrm{NU}0j}, u_{\mathrm{NA}0j}) & \sigma^2_{u_{\mathrm{NA}0j}} \end{bmatrix} \tag{14}$$

separately for each of the six main cells of the design. Finally, the covariance (correlation) between subjects' means in accented and unaccented conditions is also taken into account, as shown in (14). The resulting estimated coefficients are listed in Table 3.

As in the RM-ANOVA above, the interaction between accent and speaker group yields a significant effect ($\chi^2(2) = 20.5$, $p < 0.001$). The absolute accent contrast is considerably smaller for the laryngeal control speakers (53 ms) than for the esophageal (104 ms) and tracheoesophageal

Table 3
Estimated parameters (with standard error of estimate in parentheses) of multi-level modeling of data from Van Rossum et al. (2002)

	Model (13)	Model (15)	Model (15)
	$N = 580$	$N = 580$	$N = 445$
Fixed			
γ_{EU00}	527 (43)	526 (45)	531 (35)
γ_{EA00}	631 (38)	631 (40)	626 (36)
γ_{TU00}	427 (19)	414 (21)	429 (25)
γ_{TA00}	563 (26)	574 (29)	591 (29)
γ_{NU00}	346 (10)	344 (11)	333 (11)
γ_{NA00}	399 (14)	428 (21)	428 (23)
γ_{EU10}		−0.18 (3.99)	−0.10 (3.93)
γ_{EA10}		−4.16 (3.81)	−3.13 (4.51)
γ_{TU10}		−4.92 (3.30)	−3.57 (3.49)
γ_{TA10}		−7.32 (4.39)	−7.65 (5.03)
γ_{NU10}		−3.66 (1.35)	−4.31 (1.70)
γ_{NA10}		−7.29 (2.33)	−7.08 (2.94)
Random			
$\sigma^2_{u_{\mathrm{EU}0j}}$	13,984 (7968)	14,310 (8128)	6034 (4439)
$\sigma^2_{u_{\mathrm{EA}0j}}$	10,254 (6190)	11,854 (6928)	7680 (5459)
$\mathrm{cov}(u_{\mathrm{EU}0j}, u_{\mathrm{EA}0j})$	14,237 (6861)	15,271 (7346)	9235 (4638)
$\sigma^2_{u_{\mathrm{TU}0j}}$	2193 (1562)	2555 (1710)	3726 (2344)
$\sigma^2_{u_{\mathrm{TA}0j}}$	4474 (3024)	5770 (3561)	4909 (3685)
$\mathrm{cov}(u_{\mathrm{TU}0j}, u_{\mathrm{TA}0j})$	2354 (1698)	2149 (1864)	3044 (2276)
$\sigma^2_{u_{\mathrm{NU}0j}}$	183 (487)	353 (559)	166 (540)
$\sigma^2_{u_{\mathrm{NA}0j}}$	1128 (928)	2500 (1491)	2725 (1704)
$\mathrm{cov}(u_{\mathrm{NU}0j}, u_{\mathrm{NA}0j})$	1269 (614)	1851 (864)	1663 (845)
$\sigma^2_{e_{\mathrm{EA}ij}}$	28,927 (4545)	29,087 (4571)	23,618 (4518)
$\sigma^2_{e_{\mathrm{EU}ij}}$	28,434 (4468)	28,121 (4419)	29,841 (5419)
$\sigma^2_{e_{\mathrm{TA}ij}}$	12,748 (1900)	12,460 (1857)	11,402 (1955)
$\sigma^2_{e_{\mathrm{TU}ij}}$	22,479 (3351)	21,627 (3224)	24,155 (4236)
$\sigma^2_{e_{\mathrm{NA}ij}}$	8742 (1301)	8717 (1295)	8055 (1356)
$\sigma^2_{e_{\mathrm{NU}ij}}$	9260 (1380)	8466 (1262)	8598 (1443)

speakers (136 ms). This result supports the main hypothesis in this study. Note that this interaction was not found in the original qualitative analysis (Van Rossum et al., 2002), presumably due to the lower power of that analysis.

4.2. Adding a continuous predictor

Moreover, MLM is similar to Generalized Linear Modeling (GLM), in that it allows us to combine discrete factors such as speaker group, and continuous predictors such as peak intensity, into a single analysis. (In fact, GLM can be regarded as a specific constrained variant of multilevel modeling.) Different compensation strategies in the three speaker groups can be further investigated therefore, by including other acoustic measurements as predictors of word duration. For example, if speakers compensate between using duration cues and using intensity cues to signal accent, then this would yield a negative correlation between a word's peak intensity in dB (predictor) and its duration in ms (dependent). To investigate this type of linear effects between acoustic measurements, peak intensity was centralized to its grand mean, and then included in the model. This final model is given in (15).

$$\begin{aligned} D_{ij} = {} & \mathrm{EU}_{ij}[\gamma_{\mathrm{EU00}} + \gamma_{\mathrm{EU10}}\mathrm{Int}_{ij} + (u_{\mathrm{EU0}j} + e_{\mathrm{EU}ij})] \\ & + \mathrm{EA}_{ij}[\gamma_{\mathrm{EA00}} + \gamma_{\mathrm{EA10}}\mathrm{Int}_{ij} + (u_{\mathrm{EA0}j} + e_{\mathrm{EA}ij})] \\ & + \mathrm{TU}_{ij}[\gamma_{\mathrm{TU00}} + \gamma_{\mathrm{TU10}}\mathrm{Int}_{ij} + (u_{\mathrm{TU0}j} + e_{\mathrm{TU}ij})] \\ & + \mathrm{TA}_{ij}[\gamma_{\mathrm{TA00}} + \gamma_{\mathrm{TA10}}\mathrm{Int}_{ij} + (u_{\mathrm{TA0}j} + e_{\mathrm{TA}ij})] \\ & + \mathrm{NU}_{ij}[\gamma_{\mathrm{NU00}} + \gamma_{\mathrm{NU10}}\mathrm{Int}_{ij} + (u_{\mathrm{NU0}j} + e_{\mathrm{NU}ij})] \\ & + \mathrm{NA}_{ij}[\gamma_{\mathrm{NA00}} + \gamma_{\mathrm{NA10}}\mathrm{Int}_{ij} + (u_{\mathrm{NA0}j} + e_{\mathrm{NA}ij})] \end{aligned} \tag{15}$$

The resulting estimated parameters are also given in Table 3. In this model, the interaction of accent and speaker group does *not* reach significance ($\chi^2(2) = 5.64$, $p = 0.060$). The absence of interaction is probably due to the inclusion of peak intensity, which constitutes a stronger predictor for the word duration data (see below). Accent contrasts for the three speaker groups are of similar magnitude, and all are significant (esophageal $\chi^2 = 37.79$, $p < 0.001$; tracheoesophageal $\chi^2 = 29.62$, $p < 0.001$; laryngeal control $\chi^2 = 38.24$, $p < 0.001$). This absence of interaction falsifies the main hypothesis in this study, although it is in accordance with the original qualitative analysis (Van Rossum et al., 2002).

The present model also includes regression coefficients for the effects of peak intensity in each cell. The estimated coefficients are given in Table 3. These coefficients show interesting effects which have remained invisible in the RM-ANOVA reported above, and in the qualitative analysis in the original paper. For the laryngeal control speakers, the coefficients are negative and significant ($\gamma_{\mathrm{NU10}} = -3.7$; $\gamma_{\mathrm{NA10}} = -7.3$). This means that the word duration is shorter as the peak intensity of the word is higher, and vice versa. Laryngeal speakers signal accent by lengthening the accented word, or by increasing its intensity, but these two effects do not happen simultaneously.

For the alaryngeal speakers, however, the regression coefficients are not significant. There is no clear relation between duration and intensity for these speakers. Hence alaryngeal speakers deviate in their production of prosodic cues for accent, in that they fail to compensate between duration and intensity, as normal laryngeal speakers do. This is indeed an important conclusion about the phonetic behavior of alaryngeal speakers—and it has required MLM to come to this conclusion. This finding is relevant for further work in clinical research, and it should be applied in speech therapy aimed at improving prosody in alaryngeal speech.

Coming back to the main hypothesis, these results show that the predicted group-by-accent interaction is only observed if intensity is *not* included in the model (13), and that this interaction disappears if intensity is included (15). In the latter model, however, a similar group-by-compensation interaction is observed. Taken together, these results suggest that the latter group-by-compensation interaction describes the relevant pattern in these observed word duration data at least as well as the former group-by-accent interaction. It depends on the actual research questions whether model (13) or (15) would provide better answers, e.g., whether the predicted group-by-accent interaction is concluded to be present or absent. But conducting and comparing both analyses, which is only possible in MLM, obviously provides the most complete and accurate insight.

4.3. Robustness against missing data

Finally, we take this opportunity to demonstrate an important advantage of multi-level modeling: its robustness against missing data. The number of observations in each cell of the design matrix does *not* need to be approximately equal (Goldstein and McDonald, 1988). In general, the standard error of an estimated coefficient increases as the number of observations on which it is based decreases. Both estimated means and estimated variances follow this trend. In the data set from the study by Van Rossum et al. (2002), however, all cells contain the same number of observations. This valuable robustness of MLM is demonstrated here by randomly discarding about 1/4 of the total data set, keeping $N = 445$ out of 580 observations. Model (15) was tested again on this reduced data set; the results are also given in Table 3. All these latter estimates vary only slightly from the estimates for the full data set. Again, no group-by-accent interaction is observed ($\chi^2(2) = 4.15$, $p = 0.126$), and accent contrasts are similar to those reported above (esophageal $\chi^2 = 20.40$, $p < 0.001$; tracheoesophageal $\chi^2 =$ 29.97, $p <$ 0.001; laryngeal control $\chi^2 = 27.93$, $p < 0.001$). Note that this robustness is only observed if data are missing in a *random* fashion. If observations were predominantly missing for certain participants and/or under certain treatments, then the full and reduced data sets would not have yielded similar estimates. In short, estimating coefficients in MLM is quite robust against randomly discarding a considerable part of the data set.

5. Monte Carlo simulations

In two examples presented above, analyses with MLM yielded a significant main effect or interaction that was not reported by RM-ANOVA. This greater power in detecting effects is due to the more accurate modeling of the variance–covariance matrix, or matrices, at each level of the sampling hierarchy. This reduces the standard errors of the estimated variance components, which in turn leads to more sensitive testing of fixed effects and contrasts. Hence, it seems that MLM has more power to reject H_0 as compared to RM-ANOVA.

In order to verify this difference in power, the probabilities of rejecting H_0 in MLM and in RM-ANOVA were compared by means of Monte Carlo simulation. To this end, many fictitious data sets were generated, with a known effect size. Each set consists of data from a fictitious within-subjects experiment, using 24 items or occasions, under three treatment conditions A and B and C. One could imagine that these data are from a phonetic study in which an articulatory parameter is measured in 24 repeated occasions of the subject speaking the same phrase; the data might also conceivably come from a psycholinguistic study in which a perceptual measure is obtained for 24 word items in each condition (using the same items in each condition). From a sampling perspective, it is irrelevant whether the random within-subject variation is caused by multiple occasions or by multiple items, although a larger amount of within-subject variation is expected in the latter case. However, the appropriate variance–covariance matrix at level-1 depends on the actual design, as illustrated in model (10).

In all simulations reported here, the three treatment conditions were $\{-0.2, 0, +0.2\}$, with total variance $s^2 = 1$ and $N = 24$ participants, yielding an effect size[7] of $f = 0.16$ (between ‘small’ and ‘medium’, Cohen, 1988). Two crucial properties were varied in the generated data sets: (i) the intra-class correlation, $\rho_{\mathrm{I}} = \{0.2, 0.4, 0.6, 0.8\}$, and (ii) whether or not the data were spherical, i.e. in agreement with the sphericity

[7] Simulations were also run for other effect sizes, yielding similar power profiles. Simulations are reported here only for a single representative effect size, for the sake of clarity.

assumption. [8] Initially, variances were homogeneous (equal) in the three treatment conditions, at each level. For each combination of properties, 100 data sets were generated.

All these data sets were then analyzed with MLM (using the MLn program, Rasbash et al., 2000), with multivariate RM-ANOVA, and with univariate RM-ANOVA using the Huynh–Feldt correction (using procedures GLM and MANOVA by subjects, respectively, in SPSS version 11.5). For each data set, the resulting χ^2 (from MLM) and F test statistic (from RM-ANOVA) were compared with their appropriate critical values, at $\alpha = 0.05$. Of interest for our purposes is the probability of finding a test statistic exceeding its critical value, i.e. the probability of rejecting H_0. This is identical to the statistical power of the test, since a main effect of treatment is known to be present in the simulated data sets. The power observed in these simulations is summarized in Fig. 2.

These power profiles confirm several tendencies mentioned above. Most importantly, MLM has higher power than either univariate or multivariate RM-ANOVA in reporting a significant main effect. Presumably, this is due to its more accurate modeling of the variance–covariance matrix. If the matrix exhibits homogeneous variance as well as sphericity (left panel), then there is no difference in power between MLM and RM-ANOVA.

Second, violating the sphericity assumption (right panel) greatly reduces the power of any statistical test. The smallest reduction in power is observed for MLM, yielding a larger relative advantage of MLM in statistical power for data violating the sphericity assumption. Again, this is due to the more accurate modeling of the variance–covariance matrix in MLM.

Third, both MLM and RM-ANOVA take the intra-class correlation into account. If the intra-class correlation is large, then most of the variance is found between subjects, and hence less variance can be ascribed to the within-subject main (fixed) effect. All analyses capture this trend correctly, as a lower probability of rejecting H_0.

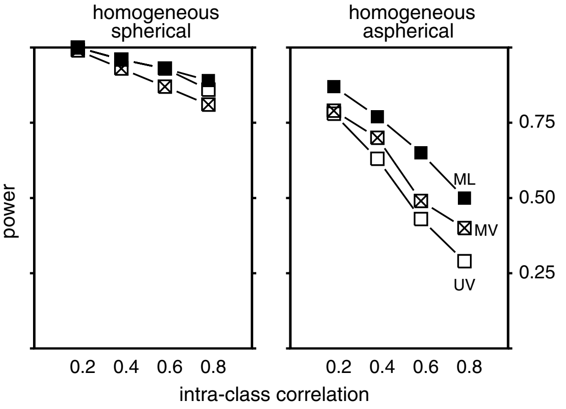

Fig. 2. Observed power in Monte Carlo simulations, using spherical data sets (left panel) and aspherical data sets (right panel), in multi-level modeling (ML, filled symbols), multivariate RM-ANOVA (MV, crossed symbols) and univariate RM-ANOVA (UV, open symbols). Data sets have homogeneous variance across treatment conditions.

Finally, among the RM-ANOVA methods, the univariate approach has higher power if the sphericity assumption holds (left panel), whereas otherwise the multivariate approach has the higher power (right panel), as previously reported (e.g. O'Brien and Kaiser, 1985; Max and Onghena, 1999).

In the simulations so far, variances were equal in the two treatment conditions. In practice, however, variances tend to differ among treatment conditions. For example, variances tend to increase with higher mean scores (e.g. Max and Onghena, 1999). Such violations of homogeneity of variance were also investigated here, by introducing an additional statistical property in the Monte Carlo simulations. New data sets were generated and analyzed, in which the variance ratios for the treatment conditions A, B, and C were specified as 1:2:3, both at level-1 and level-2. Otherwise the Monte Carlo simulations were identical to those described above. Results are summarized in Fig. 3.

The power values observed in these latter simulations with (heteroschedastic) data sets having

[8] Spherical data were generated by specifying a correlation of 0.8 for all pairs of treatments. Aspherical data were generated by specifying a correlation of 0.8 for one pair, and 0.2 for the two other pairs of treatments. Data sets were simulated using the MLn program (Rasbash et al., 2000); more details are available online at the electronic appendix to this tutorial, at URL http://www.let.uu.nl/~Hugo.Quene/personal/multilevel.

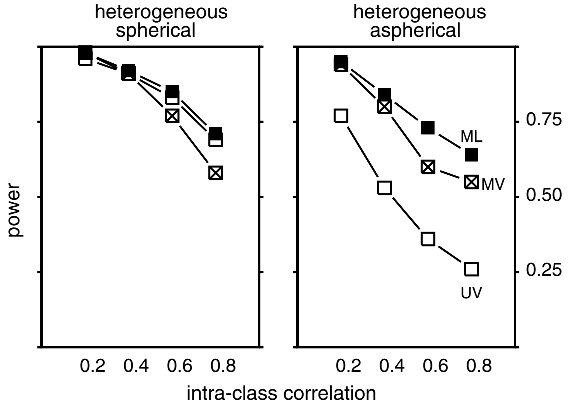

Fig. 3. Observed power in Monte Carlo simulations, using spherical data sets (left panel) and aspherical data sets (right panel), in multi-level modeling (ML, filled symbols), multivariate RM-ANOVA (MV, crossed symbols) and univariate RM-ANOVA (UV, open symbols). Data sets have heterogeneous variances across treatment conditions.

heterogeneous variances are similar to the former for (homoschedastic) data sets having homogeneous variances. The most interesting pattern is observed for heteroschedastic data sets that also violate the sphericity assumption (Fig. 3, right panel). Note that this panel corresponds best to real-life situations: between-subject and within-subject variances are unequal among treatment conditions, and pairwise differences among conditions have unequal variances. Under these most common circumstances, then, MLM yields the highest power—and even more so if there is a stronger intra-class correlation. Remarkably, the absolute power of MLM is even higher here than for homoschedastic data sets (Fig. 2, right). Apparently the non-homogeneous variance allows for more accurate estimates of the random variance components at both levels, resulting in more sensitive hypothesis testing. Hence, these simulations suggest that MLM has higher power to reject H_0 as compared to RM-ANOVA, and that this power advantage is particularly large under realistic circumstances.

Low power yields inconsistent results. If a main effect is indeed present in the population, then some studies with low power will find that there is such an effect, whereas others fail to confirm this finding. This inconsistent pattern may in turn trigger follow-up research aimed at identifying which unknown factors modulate the main effect of interest. In fact, however, the inconsistent results may be due to *low power*, and not to any modulating factors. Hence the effort expended in the follow-up research may well be inappropriate.

In summary, the Monte Carlo simulations confirm our suspicion that the statistical power is insufficient in many studies (at least those based on RM-ANOVA or similar techniques), given that effect sizes are often quite small. This yields an inconsistent pattern of results, as outlined above. The use of MLM is advocated here, as a general remedy to increase the power in hypothesis testing. If RM-ANOVA is used, however, then researchers should perform and report a post hoc power analysis (Cohen, 1988, 1992) before embarking on any follow-up research.

6. Discussion and conclusion

The two analyses and the simulations in this tutorial have attempted to demonstrate several important advantages of MLM in comparison with other analysis tools. First, MLM has higher power in finding effects and contrasts in the data. Second, there is no need for disputable assumptions, notably those of homoschedasticity (homogeneity of variance), and of sphericity. Variance and covariance components are estimated from the data, rather than postulated a priori. These variance estimates may in turn become quantities to be modeled and understood.

One might wonder, of course, whether investigators would be interested at all in random variability between subjects or between trials within subjects. After all, most studies are designed to investigate fixed effects, not random variability. Multi-level modeling might be dismissed as a sophisticated novelty, which is only useful in certain specific circumstances. However, the variances and covariances often tell their own tale about the research questions of interest. To illustrate this point, let us look again at the random coefficients in Table 3. For normal speakers, the within-speaker variances are far larger than the between-speaker variances. Within speakers, there is a lot of variability in the word durations they produce (presumably due to factors such as stress, accent, etc.), but on average speakers are quite similar: there is little variability among normal speakers' average word durations. For alaryngeal speakers, the situation is reversed. There is little variability within speakers (presumably due to each person's limited physiological capability to vary speaking rate and vowel durations), but speakers vary considerably in their average word durations (presumably due to individual differences in their "fixed" speaking rate). These interesting differences in variance ratios are immediately relevant for the research question, and for clinical practice.

The third advantage of MLM is that it takes into account the random variances at multiple levels, which are caused by *nested random factors* in a repeated measures design, and it does so simultaneously in a single full analysis. This property alone makes MLM a valuable research tool. This hierarchical modeling in turn leads to better estimates for error variances. In turn, this leads to increased power in testing the main hypotheses of interest.

Fourth, the allowance for unbalanced data across cells makes MLM very suitable for analyzing incomplete experimental designs, such as Latin-squares. Many experiments require counterbalancing of subjects across treatment conditions, to neutralize strategic effects or learning effects. Data from incomplete designs can be analyzed without any difficulty or special provisions for missing cells.

Multi-level analysis requires less stringent assumptions than conventional analysis techniques, in particular for repeated measures designs with more than one random factor. The technique can be regarded as a general form of classical ANOVA, also subsuming Generalisability Theory (Cronbach et al., 1972) and GLM. But because it requires less stringent assumptions than these analyses, it enables researchers to model their empirical data in a more insightful way. In turn, this improves scientific inference from a limited set of observations.

Of course, each statistical analysis has its advantages and disadvantages. The former were emphasized in the preceding sections. One drawback of multi-level analysis, relative to other techniques, is its computational complexity and opacity. While a simple ANOVA for a small data set can be manually computed, and verified, this is quite cumbersome for multi-level modeling. Computer software is essential to perform the analysis. Such software for MLM has become widely available in recent years, either as stand-alone programs, such as `HLM` (Bryk et al., 2001; Raudenbush and Bryk, 2002) and `MLwiN` (Rasbash et al., 2000), or as components in statistical packages, such as `lme` and `nlme` in `S-Plus` (Pinheiro and Bates, 2000) and `Proc Mixed` in `SAS` (Singer, 1998). There are considerable differences among these programs, especially in their technical details (constraints on multi-level models, default settings, estimation options, etc.). A detailed review of these programs would be outside the scope of this tutorial. The interested reader is referred to Kreft and De Leeuw (1998), Snijders and Bosker (1999) for a discussion and appraisal of the available software for multi-level analysis. As an illustration of how to use one such program, a detailed and annotated logfile is available at the electronic appendix to this tutorial, at URL http://www.let.uu.nl/~Hugo.Quene/personal/multilevel. The logfile illustrates the multi-level analyses presented above of a fictitious study into lip displacement, using the `MLwiN` program.

A second drawback is that researchers in our field still seem to be unfamiliar with this technique. A well-known conventional technique, and sophisticated modifications thereof, may therefore seem more attractive than adopting a new one. But that unfamiliarity should not last longer than necessary. Other fields of study have already profited greatly from applying MLM to their research questions, as indicated above. The new multi-level perspective provides increased insight in empirical results, and in the cognitive behavior underlying those results (e.g. Van den Bergh and Rijlaarsdam, 1996, and our analysis above). In conclusion, adopting multi-level modeling is likely to be beneficial for research in speech communication.

Acknowledgements

Our sincere thanks are due to Maya van Rossum, for providing data for re-analysis, and for helpful discussions. We also thank Sieb Nooteboom, Guus de Krom, Anne Cutler, Frank Wijnen, Saskia te Riele, Jeroen Raaijmakers, Brian McElree, Peter Dixon, and Robert L. Greene, for valuable discussions, comments and suggestions.

References

Agrawal, A.F., Brodie, E.D., Wade, M.J., 2001. On indirect genetic effects in structured populations. Amer. Nat. 158 (3), 308–323.

Beacon, H.J., Thompson, S.G., 1996. Multi-level models for repeated measurement data: application to quality of life data in clinical trials. Statist. Med. 15, 2717–2732.

Broekkamp, H., Van Hout-Wolters, B.H., Rijlaarsdam, G., Van den Bergh, H., 2002. Importance in instructional text: teachers' and students' perception of task demands. J. Educ. Psychol. 94 (2), 260–271.

Bryk, A., Raudenbusch, S., Congdon, R., 2001. HLM: hierarchical linear and nonlinear modeling. Computer program. Available from <http://www.ssicentral.com/hlm/hlm.htm>.

Bryk, A.S., Raudenbush, S.W., 1992. Hierarchical Linear Models: Applications and Data Analysis Methods. Sage, Newbury Park, CA.

Carvajal, S.C., Baumler, E., Harrist, R.B., Parcel, G.S., 2001. Multilevel models and unbiased tests for group based interventions: examples from the safer choices study. Multivar. Behav. Res. 36 (2), 185–205.

Cochran, W., 1977. Sampling Techniques, third ed. Wiley, New York.

Cohen, J., 1988. Statistical Power Analysis for the Behavioral Sciences, second ed. Lawrence Erlbaum Associates, Hillsdale, NJ.

Cohen, J., 1992. A power primer. Psychological Bull. 112 (1), 155–159.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N., 1972. The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. Wiley, New York.

Goldstein, H., 1991. Nonlinear multilevel models with an application to discrete response data. Biometrika 78, 42–51.

Goldstein, H., 1995. Multilevel Statistical Models, second ed. Edward Arnold, London.

Goldstein, H., McDonald, R., 1988. A general model for the analysis of multilevel data. Psychometrika 53, 455–467.

Goldstein, H., Healy, M.J., Rasbash, J., 1994. Multilevel time series models with applications to repeated measures data. Statist. Med. 13, 1643–1655.

Haggard, E.A., 1958. Intraclass Correlation and the Analysis of Variance. Dryden Press, New York.

Hall, D.B., Bailey, R.L., 2001. Modeling and prediction of forest growth variables based on multilevel nonlinear mixed models. Forest Sci. 47 (3), 311–321.

Hox, J., 1995. Applied Multilevel Analysis. TT Publicaties, Amsterdam.

Kirk, R., 1995. Experimental Design: Procedures for the Behavioral Sciences, third ed. Brooks/Cole, Pacific Grove, CA.

Kish, L., 1967. Survey Sampling, second ed. Wiley, New York.

Kreft, I.G., De Leeuw, J., 1998. Introducing Multilevel Modeling. Sage, London.

Lochner, K., Pamuk, E., Makuc, D., Kennedy, B.P., Kawachi, I., 2001. State-level income inequality and individual mortality risk: a prospective, multilevel study. Amer. J. Public Health 91 (3), 385–391.

Longford, N., 1993. Random Coefficient Models. Oxford University Press, New York.

Max, L., Onghena, P., 1999. Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. J. Speech Language Hearing Res. 42, 261–270.

Maxwell, S.E., Delaney, H.D., 2004. Designing Experiments and Analyzing Data: A Model Comparison Perspective, second ed. Lawrence Erlbaum Associates, Mahwah, NJ.

McCulloch, C.E., Searle, S.R., 2001. Generalized, Linear, and Mixed Models. Wiley, New York.

Merlo, J., Ostergren, P.O., Hagberg, O., Lindstrom, M., Lindgren, A., Melander, A., Rastam, L., Berglund, G., 2001. Diastolic blood pressure and area of residence: multilevel versus ecological analysis of social inequity. J. Epidemiol. Community Health 55 (11), 791–798.

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

Pedhazur, E., Schmelkin Pedhazur, L., 1991. Measurement, Design, and Analysis: An Integrated Approach. Lawrence Erlbaum Associates, Hillsdale, NJ.

Pinheiro, J.C., Bates, D.M., 2000. Mixed-Effects Models in S and S-Plus. Springer, New York.

Raaijmakers, J.G., Schrijnemakers, J.M., Gremmen, F., 1999. How to deal with "the language-as-fixed-effect fallacy": common misconceptions and alternative solutions. J. Memory Language 41, 416–426.

Rasbash, J., Browne, W., Goldstein, H., Yang, M., Plewis, I., Healy, M., Woodhouse, G., Draper, D., Langford, I., Lewis, T., 3 December 2000. A user's guide to MLwiN. Computer program. Available from <http://multilevel.ioe.ac.uk/>.

Raudenbush, S.W., Bryk, A.S., 2002. Hierarchical Linear Models: Applications and Data Analysis Methods, second ed. Sage, Thousand Oaks, CA.

Reise, S.P., Duan, N.H., 2001. Introduction to the special issue on multilevel models. Multivar. Behav. Res. 36 (2), 153.

Searle, S., 1987. Linear Models for Unbalanced Data. Wiley, New York.

Searle, S.R., Casella, G., McCulloch, C.E., 1992. Variance Components. Wiley, New York.

Singer, J., 1998. Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. J. Educ. Behav. Statist. 23, 323–355.

Sluijter, A.M.C., 1995. Phonetic Correlates of Stress and Accent. Foris, Dordrecht.

Snijders, T., Bosker, R., 1999. Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling. Sage, London.

Van den Bergh, H., Rijlaarsdam, G., 1996. The analysis of writing process data: a mini longitudinal study. In: Levy, C.M., Ransdell, S. (Eds.), The Science of Writing: Theories, Methods, Individual Differences, and Applications. Lawrence Erlbaum, Mahwah, NJ, pp. 207–232.

Van der Leeden, R., 1998. Multilevel analysis of repeated measures data. Quality and Quantity 32, 15–19.

Van Rossum, M., De Krom, G., Nooteboom, S., Quené, H., 2002. 'Pitch' accent in alaryngeal speech. J. Language Speech Hearing Res. 45, 1106–1118.

Winer, B., 1971. Statistical Principles in Experimental Design, second ed. McGraw-Hill, New York.