NISS

Questioning Multilevel Models

Jan de Leeuw and Ita G. G. Kreft

Technical Report Number 31
December, 1994

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

QUESTIONING MULTILEVEL MODELS

JAN DE LEEUW AND ITA G.G. KREFT

November 26, 1994

1. INTRODUCTION

In this paper[1] we discuss some of the practical problems in using multilevel techniques, by looking into the choices users of these techniques have to make. It is difficult, of course, to define "user". Different users have different degrees of statistical background, computer literacy, experience, and so on. We adopt a particular operational definition of a "user" in this paper, which certainly does not apply to all users. Our "user" is defined by the set of questions asked by the Statistical Standards and Methodology Division of the National Center for Education Statistics (as formulated in a letter of 9-16-93 of Bob Burton of NCES to Jerry Sacks of the National Institute of Statistical Sciences) whose purpose is to evaluate the practical usefulness of multilevel modeling for educational statistics. We cannot discuss, let alone answer, all the questions from NCES in this paper. Many of them will require additional statistical and computational research, but they illustrate nicely some of the practical methodological problems in using hierarchical linear models. They also illustrate the dominant position of the terminology and notation of Bryk and Raudenbush [7], and of the computer program HLM [8] in the field of "official" educational statistics. In many cases it seems as if "fitting a multilevel model" and "using HLM" are

Key words and phrases. Variance components, hierarchical models, mixed models, multilevel models, random coefficients.

[1]This version of the paper has benefited greatly from written and spoken comments by John Tukey. The remaining errors, mistakes, and ambiguities are all ours. A version of this paper was presented in October 1993 at a workshop on *Hierarchical Linear Models: Problems and Prospects*, at the RAND Corporation in Santa Monica. Another, considerably shorter, version of the paper will be published in the *Journal of Educational and Behavioral Statistics*, Summer 1995, which contains the proceedings of the RAND workshop. Research of the first author supported through the National Sciences Foundation Grant No. DMS-9208758.

seen as identical activities. They are not, of course. There is more to structural equations modeling than LISREL, and there is more to Aspirin than Bayer. To avoid confusion, we shall not use the term "hierarchical linear models," and if we say HLM we mean the computer program of that name.

We shall keep the statistics and mathematics as simple as possible. We shall also concentrate on the situation in which we have a relatively large number of relatively small groups. The situation in which we have only two or three groups does not really interest us here, and the situation in which we have a large number of very small groups (twins, couples) also requires a slightly different emphasis. We are thinking in terms of at least 20, but maybe as much as 1000, groups of size at least 5, but maybe as large as 50. This seems to cover most studies in which the individuals are students and the groups are schools or classes.

Throughout, we shall try to keep the following quotation from Wilk and Kempthorne [66] in mind.

> We feel that any mathematical assumptions employed in the analysis of natural phenomena must have an explicit, recognizable, relationship to the physical situation. In particular, if the analysis of variance is to be generally useful in the interpretation of experimental data it is necessary that its meaning and justification should transcend the set of arbitrary assumptions which are usually put forth.

It is likely that papers looking critically at multilevel models fall into two classes, which try to answer two different questions, aptly described by Searle [54].

> This approach answers the question "given this model and its definitions, what consequences can I deduce ?" However, this is not the prime question for the biologist. His concern is usually "given these data and their origin, what model do I use ?"

We shall try to concentrate on the second type of question, although lapses into the first mode are unavoidable. The NCES questions will be discussed in terms of a number of *choices* the user has to make. Here is a brief list. The user has to choose

- a selection and coding of her variables;
- a model from the class of regression models;
- a loss function to measure goodness-of-fit;
- an algorithm to minimize the loss function;
- a computer program to implement the algorithm.

We shall see that all these choices are nontrivial, but our discussion

will mainly emphasize the choice of the model, the loss function, and the technique. And, of course, the consequences of these choices.

2. Musings on Linear Regression

The first, rather general, question in the list posed by NCES is

> Is some form of hierarchical linear model always preferable when conducting analysis with independent variables from two levels of a hierarchical data set ? Are there alternatives to the HLM software that NCES should consider using ?

We shall postpone the question about software to a later section, and extensively comment on the problem of model choice here. From the modeling point of view, the first important choice is to decide what is *random* and what is *fixed* in our regression models. Predictors can be fixed and random, and coefficients can be fixed and random. Although this choice in itself may seem somewhat esoteric, it has major consequences for subsequent computations, although perhaps not always for the outcome of these computations. Before we discuss the choices, we obviously have to discuss the alternatives the user can choose from.

Let us start our discussion with the usual linear regression model. So we have n individuals and p predictors. The outcomes are in an $n-$element vector $y = \{y_i\}$, the values on the predictors in an $n \times p$ matrix $X = \{x_{is}\}$. We suppose

$$\underline{y} = X\beta + \underline{\epsilon}. \tag{1}$$

Some aspects of (1) are of importance, and since they usually are discussed rather vaguely we emphasize them here. In the first place we distinguish random variables from fixed quantities by underlining them [32]. In discussion this class of regression models, the distinction between what is fixed and what is random is crucial, and the underlining notation helps to emphasize the difference between the two. Of course what the distinction actually *means* in a practical data analysis context has been the subject of much debate, because it is intimately linked with the foundations of statistics and with the quarrel between the Bayesian and Frequentist schools.

We use frequentist terminology in this paper. Thus we think of (1) as a model that describes a hypothetical sequence of replications of the experiment that generated the data. Our statistical model does not describe the outcome of a single experiment, or of an actual sequence of replications, but it models a hypothetical sequence of replications. In this hypothetical sequence X remains fixed, i.e. it is exactly the

same in each replication. The coefficients β are also fixed over replications, but we do not know what their values are. They are *parameters* that have to be *estimated.* The *disturbances* $\underline{\epsilon}$ are different for each of the hypothetical replications, and they vary according to specified probability distributions. In particular, we assume, for their expected value and dispersion matrix,

$$\mathbf{E}(\underline{\epsilon}) = 0, \tag{2}$$

and

$$\mathbf{V}(\underline{\epsilon}) = \sigma^2\mathcal{I}. \tag{3}$$

The model (1)-(3) says, in essence, that the disturbances are uncorrelated. They all have the same expectation and variance, i.e. they are not systematically related to X. Or to anything else within the model, for that matter.

Some general comments are in order here. In the first place the model (1)-(3) was really meant for situations in which the predictors in X are under experimental control, and can be assumed to be measured without error. That is, they are meant for designed experiments. The x_{is} are fixed quantities, i.e. they remain the same over the hypothetical replications, which means that in order to use the model we must have have a way of physically keeping them the same. This does not happen very often in educational statistics. If we regress school success on IQ, we are usually not interested in replications in which the individual has the same IQ all the time, only different school success. Both variables covary, i.e. it looks as if we should use a model with a random predictor. Fortunately, this problem can be solved quite easily. We assume that (1) models the conditional distribution of $\underline{y}_i$ given $\underline{x}_i = x_i$, and, if we want to, we can model the marginal distribution of $\underline{x}_i$ separately to get a model for the joint distribution of $(\underline{y}_i, \underline{x}_i)$. In this, we agree with Beran and Hall [5], page 1971:

> We should comment on our decision to condition on the variables x_i. In our view, regression is intrinsically the study of functional relationships where the design variables are held fixed, that is, are regarded as nonrandom. If the $x_i's$ are not conditioned upon, then the study is one of correlation, not regression.

A second problem that remains is that assuming linearity and homoscedasticity of the regression in the joint distribution is a very strong assumption, which is unlikely to be even approximately true. It forces us to take a more modest approach, in which models are used to generate tools for compact *description* and/or tools for *prediction.* We do

not have to worry about the model being true (it obviously is not), we only have to worry if the procedures it generates do their job of summarizing the information in the data and extrapolating into the future well enough. We think it is still the general consensus that the procedures usually generated by the linear regression model (1)-(3) do very well, given how strong and unrealistic the model is. It is still the workhorse of applied statistics, in fact it sometimes seems as if applied statistics *is* linear regression analysis.

This points to a third general point, which is of tremendous importance, and which is not often discussed. Statistical models are *languages* that users in a particular field have to learn, and that they use to talk to each other efficiently. Regression analysis, path analysis, factor analysis, survival analysis are all examples of this. There is a tendency to narrow down the language even more, so that for example in the seventies LISREL became the language of choice for a large group of scientists in various disciplines. In educational statistics the multilevel framework provides a language that encompasses and supersedes the older language of contextual analysis, and there seems to be a tendency to narrow it down even more to the language of the HLM program. But this means that in the field it becomes difficult to talk about hierarchical data structures without adopting the terminology (and constraints) of the HLM program. The NCES questions seem to indicate that it is not clear to everybody in educational statistics that the current multilevel language, or its HLM dialect, should really occupy this exclusive position.

3. On Random Coefficients

Finally, another assumption which is implicit in (1) is that the regression coefficients β are the same for all individuals. Starting with Wald in 1947, economists have been critical about this assumption too, although as usual for the wrong reasons. In his textbook, page 216, Klein [38] says

> Individuals differ greatly in behavior, and it may not be possible to obtain observations on a sufficiently large number of variables so that each unit may be considered to behave according to the same structural equation. We are then faced with the problem of interpreting a single estimated equation as representative in some sense of a large number of underlying equations.

The quotation is interesting, because it states explicitly that we need multiple equations because we do not have enough predictors. If we

had all relevant predictors in our study, we could use a single equation for all individuals, but since this is impossible, or at least impractical, the equations will vary around some average equation. It does assume, however, that there is some "true" model, which we can only approximate imperfectly because we will never have enough individuals to estimate its parameters. This is, of course, Platonic Idealism. Nevertheless the notion that individuals have their own regression coefficients, and that these do not vary too wildly, seems useful.

We formalize this by using the notion of random coefficients. The model is

$$\underline{y}_i = x_i'\underline{\beta}_i + \underline{\epsilon}_i, \tag{4}$$

$$\underline{\beta}_i = \beta + \underline{\delta}_i. \tag{5}$$

where $\underline{\delta}_i$ are independent and identically distributed with zero expected value and dispersion Ω. Moreover they are independent of the $\underline{\epsilon}_i$. Thus $\underline{y}$ has expectation $X\beta$, as in the fixed coefficient model, but we now have heteroscedasticity because

$$\mathbf{V}(\underline{y}_i) = x_i'\Omega x_i + \sigma^2. \tag{6}$$

Thus the variance of $\underline{y}_i$ is a linear function of the squared length of x_i, in the metric Ω. We should be able to see such an effect in residual plots.

Once again we emphasize that the distinction between fixed and random coefficients is important, because it changes the definition of the population over which we generalize. If we repeat our experiment, then we do not expect individual i to have the same regression coefficients in each hypothetical replication. The regression coefficients vary, both within individuals and between individuals, around a population mean. Some applications of these "models of the second kind in regression analysis" are discussed by Fisk [25].

4. Regression in Multiple Populations

We now analyze the more general situation in which there are m groups, indexed by j. A straightforward generalization of (1) is

$$\underline{y}_j = X_j\beta_j + \underline{\epsilon}_j. \tag{7}$$

Now the $\underline{y}_j$ and the $\underline{\epsilon}_j$ are vectors of length n_j, the number of individuals in group j. Matrix X_j is $n_j \times m$. It conceivably makes sense also to assume

$$\mathbf{E}(\underline{\epsilon}_j) = 0, \tag{8}$$

and

$$\mathbf{V}(\underline{\epsilon}_j) = \sigma_j^2 \mathcal{I}. \tag{9}$$

Finally we assume the different $\underline{\epsilon}_j$ are independent of each other.

There is nothing wrong with model (7). It merely says that the same regressors apply to each of the groups. But if we fit the model, we can fit it to each of the m groups separately, because none of the parameters are common to the groups. Especially if there is a large number of relatively small groups, for instance students from many school classes, where each class has somewhere around $10 - 20$ students, this is not very attractive. We ignore the fact that schools are all part of the same system, and that consequently the regressions are likely to have something in common. We would like to incorporate this communality into the model. One way to do this is requiring that some of the parameters are equal in all groups. There are two obvious choices

$$\beta_1 = \beta_2 = \cdots = \beta_m, \tag{10}$$

$$\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_m^2. \tag{11}$$

But this is a clear case of throwing the baby out with the bath water. Although schools are related, and must have something *in common*, we do not want to assume they are *identical*. There are a number of ways out of this dilemma. We discuss three of them in this section. The first one assumes partial parameter identity. The second solution simply fits model (7), and then tests if the additional specifications in (10)-(11) are true. The third one uses a random coefficient model that takes the hierarchical structure of students in schools into account.

The Analysis of Covariance is an example of the first approach. We assume (11), and we assume that all slopes, but not all intercepts, are equal. More generally, we can require $\beta_j = Z_j\gamma$, where the Z_j are chosen in such a way that some elements of the β_j are equal. In ANCOVA, for instance, we have mp parameters β_{sj}, and we replace them by $p - 1$ slopes and m intercepts. If the slopes are in a vector β and the intercepts in a vector α, we can set

$$\beta_j = Z_j\gamma = \begin{pmatrix} e_j' & 0 \\ 0 & \mathcal{I} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{12}$$

with e_j the j-th unit vector. Thus we see that the partial identity approach leads to linear constraints on the β_j. Combining these constraints with (7) simply gives the fixed-effects linear model

$$\underline{y}_j = X_j Z_j \gamma + \underline{\epsilon}_j, \tag{13}$$

which can be fitted with ordinary least squares methods.

To illustrate the second approach, suppose we compute, with any one of our trusted OLS programs, estimates $\hat{\beta}_j$ and $\hat{\sigma}_j^2$. Then we know that, under normality, and assuming (10)-(11),

$$\hat{\beta}_j \sim \mathcal{N}\Big\{\beta, \sigma^2(X_j'X_j)^{-1}\Big\}. \tag{14}$$

Thus it is simple to test equality, because (14) is just a simple linear model for the coefficients. In the same way we can test equality of the σ_j^2, which are independent chi-squares under normality. These tests correspond nicely with a partitioning of the normal log-likelihood into within-group and between-group regressions. What we see from this analysis is that from the data analysis point of view this *two-step procedure*, which first fits a model for each group separately, and then analyzes the parameter estimates in a second step is potentially much more informative then the *one-step procedure,* which just plugs in (10)-(11) into (7)-(9) and fits the resulting model. The two-step approach allows us to partition the log-likelihood (or, equivalently, the residual sum of squares), and it allows us to look at residuals at two levels. Of course if there are many regressors, and the groups are small, the within-group regression coefficients will be unstable, or even unidentified.

In random coefficient models, we go a slightly different route. We take a position which is in between "separate models for all schools" and "complete equality". In an excellent review paper Spjøtvoll [62] discusses an example in which length and thickness of a number of cucumbers at five points in time are compared.

> It is seen that the points for each cucumber almost lie on a straight line. Hence a straight line can be used to represent the relationship between thickness and length for a given cucumber. But each cucumber seems to have its own line. The cucumbers are chosen at random from a large number of cucumbers of a certain variety. Hence the regression lines must be considered as random. The individual cucumbers can be characterized by their straight-line relationships. To characterize the whole population of cucumbers it is natural to look at the distribution of these lines. The expected values of the regression coefficients, their variances and covariances are then the parameters of interest.

To make the quotation relevant for educational statistics, just substitute "student" for "cucumber". We see the reason for using random coefficients here. If we do another hypothetical replication, we do not use the same cucumbers, but we draw a new sample and watch them

grow. More discussion of this is under the heading "Fixed or Random" in textbooks such as [55], [53].

The model becomes

$$\underline{y}_j = X_j\underline{\beta}_j + \underline{\epsilon}_j, \tag{15}$$

with

$$\underline{\beta}_j = \beta + \underline{\delta}_j. \tag{16}$$

Compare this with (4)-(5). In the earlier model we assume that each individual has her own regression coefficients, and these coefficients are independent between individuals. In (15) we assume each *group* has its own regression coefficients, which are independent over groups. But the coefficients are *identical* for different individuals in the same group. We model the coefficients as random, which means that we modify our definition of a population (our hypothetical sequence of replications). The slopes and intercepts are no longer fixed numbers, which are constant within schools and maybe even between schools, but they also vary over replications. In order to complete the specification we also assume that the *second-level disturbances* $\underline{\delta}_j$ are independent, are independent of the first-level disturbances $\underline{\epsilon}_j$, have zero expectation, and have dispersion matrix Ω. If the second level disturbances are identically equal to zero, then we are back in the situation (10)-(11).

Observe that the parameters σ_j^2 are not usually modeled as random, but are considered to be fixed, and in most cases actually the same. The paper by Aragon [2] is an exception. The only obvious reason for this is mathematical convenience, because assuming random individual level dispersions takes us out of the usual normal theory framework. Also, observe that using fixed regressors is mathematically very convenient in the random coefficient context. If both regressors and coefficients are random, then, even if they are both normal, their product is certainly not normal. This makes multilevel path analysis models [20] inherently more complicated.

By combining (15) and (16) we see that

$$\underline{y}_j = X_j\beta + X_j\underline{\delta}_j + \underline{\epsilon}_j, \tag{17}$$

which implies

$$\mathbf{V}\left(\underline{y}_j\right) = X_j\Omega X_j' + \sigma_j^2\mathcal{I}. \tag{18}$$

Individuals in the same school have correlated disturbances, and the correlation will be larger if their predictor profiles are more similar, in the metric Ω. This is an interesting consequence of the specification (15)-(16), but understanding (15)-(16) itself is clearly more basic. It

will be difficult, even for sophisticated users, to interpret the variance and covariance components in (18) directly.

5. Multilevel Data and Models

We have seen, in the previous section, that random coefficient models are a convenient compromise between separate fixed coefficient models for each group, and models with all coefficients equal for each group. They are a convenient compromise, because we expect them to give more stable estimates than separate models and more interesting parameters than equal coefficients. They are also more plausible, because we they reflect the structure in which many educational data sets are put together. Plausibility, by the way, is nice, but it cannot be the only criterion. The more parameters we add, the more plausible the model becomes, and if we continue long enough each individual has her own set of parameters which can be used to get a very good fit, that will never stand up to replication. Thus we also need Parsimony, Plausibility's eternal enemy. Statistics is the battle between Plausibility and Parsimony.

Our regression situation becomes more complicated, but also more interesting, if we have variables describing individuals (students) as well as variables describing groups (schools). Combining them in a single analysis is called *multilevel analysis*. In multilevel analysis we combine the approaches discussed in the previous section. We use the linear restrictions of the form $\beta_j = Z_j\gamma$ to reduce the number of free regression parameters, and we use the idea of random coefficients to model the idea that schools are sampled and that we cannot expect to explain all relevant variation with only a few regressors. The combined model, which replaces (15)-(16), is

$$\underline{y}_j = X_j\underline{\beta}_j + \underline{\epsilon}_j, \tag{19}$$

$$\underline{\beta}_j = Z_j\gamma + \underline{\delta}_j. \tag{20}$$

We now clearly see two different regression models on two different levels. The first level model (19) is complemented by the second level model (20).

We have now collected enough ammunition to turn to NCES question number four.

> Some analysts are more comfortable presenting HLM results in terms of a combined model, i.e., a single regression equation containing interaction terms. Others prefer to discuss the coefficients without recourse to a single regression equation. Are the two approaches equally valid ?

Let us translate this into formulas, because at least partly it is a question about formulas. If we substitute (20) into (19) we have

$$\underline{y}_j = X_j Z_j \gamma + X_j \underline{\delta}_j + \underline{\epsilon}_j. \tag{21}$$

If we look at the fixed part of (21) we see that

$$\mathbf{E}(\underline{y}_j) = X_j Z_j \gamma. \tag{22}$$

In (22) the *cross-level interactions* are formed as products of the first-level regressors x_s and the second-level regressors z_r. In a sense, there is not much to choose. The single-equation and the two-equation formulation describe the same model. From the point of view of interpretation, however, the two formulations are quite different. We feel it is very difficult, perhaps impossible, to interpret (21) without going back to (19)-(20). It is of course possible to interpret the fixed effects in (21), because we have a lot of experience with interpreting interactions in fixed effect situations. Compare the useful reviews by Aiken and West [1] and Cox [16]. It is however quite impossible to come up with a convincing interpretation of the structure of the disturbance term in (21) without referring to (19)-(20). The disturbance term in question is $X_j \underline{\delta}_j + \underline{\epsilon}_j$, and its dispersion matrix is $X_j \Omega X_j' + \sigma_j^2 \mathcal{I}$. We have seen, in the previous section, that it is difficult to make direct sense of especially the covariance components.

For completeness, the multilevel random coefficient model explained above must be distinguished from two-level models using fitted coefficients. In the "slopes-as-outcomes" approaches of Burstein and his co-authors [10] it seems that the model is (1)-(3), but with in addition

$$\hat{\beta}_j = Z_j \gamma + \underline{\delta}_j, \tag{23}$$

with

$$\hat{\beta}_j = (X_j' X_j)^{-1} X_j' \underline{y}_j, \tag{24}$$

which are just the fitted regression coefficients. These assumptions imply that $\beta_j = Z_j \gamma$, and that

$$\mathbf{V}(\underline{\delta}_j) = \sigma_j^2 (X_j' X_j)^{-1}. \tag{25}$$

As a consequence, this fitted coefficients model is identical to the model (13), i.e. a model with cross-level interactions, but no intraclass covariance component structure. The Gauss-Markov estimates are the one-step OLS estimates, and not the two-step estimates typically used in slopes-as-outcomes. A nice treatment of slopes-as-outcomes was published in 1974 by Hanushek [28]. He assumes (in a simple special case)

both (19) and (20), and then adds

$$\hat{\beta}_j = \underline{\beta}_j + \underline{\eta}_j, \tag{26}$$

where $\underline{\eta}_j$ is the sampling error of the regression coefficients, given by (25).

6. Estimates, Loss Functions, and Global Fit Measures

The one-step and the two-step model, discussed in the previous section, suggest two different ordinary least squares methods for fitting the model. This was already discussed in detail by Boyd and Iverson [6]. We follow the treatment of [19]. The two-step method first estimates the β_j by

$$\hat{\beta}_j = (X_j'X_j)^{-1}X_j'\underline{y}_j, \tag{27}$$

and then γ by

$$\hat{\gamma} = (\sum_{j=1}^m Z_j'Z_j)^{-1}\sum_{j=1}^m Z_j'\hat{\beta}_j. \tag{28}$$

The one-step method estimates γ directly from (21) as

$$\hat{\gamma} = (\sum_{j=1}^m Z_j'X_j'X_jZ_j)^{-1}\sum_{j=1}^m Z_j'X_j'\underline{y}_j = (\sum_{j=1}^m Z_j'X_j'X_jZ_j)^{-1}\sum_{j=1}^m Z_j'X_j'X_j\hat{\beta}_j. \tag{29}$$

Both methods provide unbiased estimates of γ, they are non-iterative, easy to implement, and because they are linear in the observations it is trivial to give an expression for their (unknown) dispersion matrices. Nevertheless they have fallen into disgrace, because they are neither BLUE nor BLUP [27], [49]. On the basis of the computational experience we have so far (which is quite minimal) we feel that they still deserve a fighting chance.

The next candidate that comes to mind is based on the BLUE, i.e. the best linear unbiased estimate. If we knew σ_j^2 and Ω, then we could compute the BLUE by

$$\hat{\gamma} = \Big\{\sum_{j=1}^m Z_j'X_j'(X_j\Omega X_j' + \sigma_j^2\mathcal{I})^{-1}X_jZ_j\Big\}^{-1}\sum_{j=1}^m Z_j'X_j'(X_j\Omega X_j' + \sigma_j^2\mathcal{I})^{-1}\underline{y}_j. \tag{30}$$

This looks horrible, but it can be simplified to

$$\hat{\gamma} = \Big\{\sum_{j=1}^m Z_j'W_j^{-1}Z_j\Big\}^{-1}\sum_{j=1}^m Z_j'W_j^{-1}\hat{\beta}_j, \tag{31}$$

where

$$W_j = \Omega + \sigma_j^2(X_j'X_j)^{-1}. \tag{32}$$

Observe that W_j is the dispersion of the OLS estimate $\hat{\beta}_j$. The formal similarity of (28), (29), and (31) is clear. They can all be thought of as two-step methods, which first compute the $\hat{\beta}_j$, and then do a weighted regression of the $\hat{\beta}_j$ on the Z_j. Of course (31) is useless by itself, because we do not know what σ_j^2 and Ω are, but a method to compute consistent estimates of these variance parameters from the OLS residuals is discussed in [19]. This adapts a method proposed by Swamy [60] to the multilevel model. Again, we think a more detailed comparison of these simpler methods with the complicated iterative methods such as HLM [8], or VARCL [42], or ML2 [47] would be useful. The least squares methods are computationally simpler, and easier to understand and explain. Moreover it is generally simpler to study their statistical properties. For the case in which we first estimate the variance components, the statistics are still quite complicated [35].

This also brings us to the next question asked by NCES.

> Most discussion of HLM results centers on the individual coefficients: the betas and gammas. There is of course some interest in the overall measures, such as the proportion of variance explained. What is the best way to obtain and present overall measures when using HLM ?

Each of the methods discussed above gives one way to compute the "proportion of variance explained". We have residual sums of squares in each of the two steps. We get a somewhat more integrated picture by using the *Analysis of Deviance*, which is based on the multinormal likelihood function. Let us combine fixed and random coefficient models to

$$\underline{y}_j = X_j\underline{\beta}_j + \underline{\epsilon}_j, \tag{33}$$

$$\underline{\beta}_j = \beta_j + \underline{\delta}_j. \tag{34}$$

The two additional specifications we can either impose, or test, or both, within this model are

$$\beta_j = Z_j\gamma, \tag{35}$$

and

$$\Omega = 0. \tag{36}$$

A special case of (35) is equality of the β_j, another special case is (random effects) ANCOVA. Of course (36) is the hypothesis that the regression coefficients have no random variation.

The multinormal deviance for model (33)-(36) is, ignoring the usual constants,

(37)
$$\Delta = \sum_{j=1}^m \log \mid X_j \Omega X_j' + \sigma_j^2 \mathcal{I} \mid + \sum_{j=1}^m (y_j - X_j \beta_j)' [X_j \Omega X_j' + \sigma_j^2 \mathcal{I}]^{-1} (y_j - X_j \beta_j).$$

This can be simplified by writing $y_j = X_j \hat{\beta}_j + r_j$, where $\hat{\beta}_j$ is any OLS estimate. We find that, except again for some constants,

(38)
$$\Delta = \sum_{j=1}^m \Big\{ \log \mid W_j \mid + (n_j - p) \{ \log \sigma_j^2 + \frac{\hat{\sigma}_j^2}{\sigma_j^2} \} + (\hat{\beta}_j - \beta_j)' W_j^{-1} (\hat{\beta}_j - \beta_j) \Big\}.$$

Here $\hat{\sigma}_j^2$ is the OLS estimate of the residual variance, i.e.

$$\hat{\sigma}_j^2 = \frac{r_j' r_j}{n_j - p}. \tag{39}$$

The derivation of (38) from (37) is, for example, in [19]. It seems that most "overall measures" that are "useful" are components of (38). We see the residual individual level variance σ_j^2 in each group, while the two components of W_j are the *parameter variance* Ω and the *estimation variance* $\sigma_j^2 (X_j' X_j)^{-1}$. These components are discussed extensively in [7].

If we want to establish how much variance of the $\hat{\beta}_j$ is "explained" by the Z_j we merely have to compute the matrix

$$\sum_{j=1}^m \hat{W}_j^{-1} (\hat{\beta}_j - Z_j \hat{\gamma}) (\hat{\beta}_j - Z_j \hat{\gamma})', \tag{40}$$

and look at its diagonal or trace. Here $\hat{W}_j$ and $\hat{\gamma}$ are the maximum likelihood estimates, computed by minimizing the deviance (38) over the free parameters.

In HLM, and in some of the other multilevel programs as well, the deviance that is actually minimized is defined slightly differently. Instead of minimizing the deviance of the data, we minimize the deviance of the least squares residuals. This leads to Restricted Maximum Likelihood or REML estimates [29]. In the multilevel context the relevant algebra is in the appendix of the book by Bryk and Raudenbush [7], or in the paper by De Leeuw and Liu [21]. REML estimates are generally considered to be superior to the maximum likelihood estimates

based on the deviance of the data, but the evidence of their superiority in complicated cases, and in multilevel analysis in particular, is not too convincing. The precise asymptotics for both ML and REML has been worked out [44], [17] but as usual the results are not very helpful. Careful Monte Carlo studies in simpler cases [59] do not lead to unambiguous recommendations. Clearly a great deal more of research, of the theoretical and the Monte Carlo variety, is needed here.

Another question which is of some interest from the practical point of view is how we estimate the $\underline{\beta}_j$. Obviously we can use the unbiased and consistent estimates $\hat{\beta}_j$ or $\hat{\beta}_j = Z_j\gamma$, just as we would do in the fixed coefficient case. This is not what is normally done, however. One of the key selling points of multilevel approaches is the *shrinkage estimator* which is used to *borrow strength* from the other contexts (groups, schools). In this approach we estimate β_j by using the conditional expectation (or the linear regression, in the non-normal case) of $\underline{\beta}_j$ given $\underline{y}$. The shrinkage estimate has the simple expression

$$\tilde{\beta}_j = \hat{\Theta}_j\hat{\beta}_j + (\mathcal{I} - \hat{\Theta}_j)Z_j\hat{\gamma}, \tag{41}$$

with

$$\hat{\Theta}_j = \hat{\Omega}\hat{W}_j^{-1}. \tag{42}$$

Thus the shrinkage estimator $\tilde{\beta}_j$ is in the class of "matrix weighted averages", and the algebra and geometry derived in Chamberlain and Leamer [12] apply. Using the weighted average interpretation can help in the understanding of the regression coefficients. It can also help in understanding the frustration of the principal of an excellent school, who sees the predictions of success of her students shrunken towards the mean.

The fact that we actually have three different estimates of β_j offers many opportunities for diagnostics which have not really been explored so far. In fact, the emphasis in the literature has been on the appropriateness and the plausibility of the model, and not on the ways in which it can be violated. This is perhaps a useful attitude in the initial stages of development, but the time has come to become more realistic. One possibility is to relax the assumptions and to fit more general models. As we know, going this route means going further into the minefield of Plausibility, declaring war on Parsimony and its faithful ally Stability. The other possibility is to use diagnostics, either graphical or computational. There have been a few attempts to develop tools for the mixed linear model [4], [14], [40], but their usefulness has not really been explored.

7. Algorithms and Computer Programs

Some people think, perhaps, that it is irrelevant for the ordinary user which algorithm is used to compute, say, maximum likelihood estimates. Moreover, it is equally irrelevant which computer program is used to compute the estimates. But this is true in the same sense that it is irrelevant which means of transportation you use to get to your work. Eventually you will get there alright, no matter what means of transportation you use, but walking takes hours, the bus is unpleasant, and an old car breaks down all the time. The review by Kreft, De Leeuw, and Kim [39] shows that algorithms *do* matter, and that consequently the NCES question about software makes perfect sense. Related comparisons are in Van der Leeden, Vrijburg, and De Leeuw [63]. On the basis of this comparison, the answer to

> Are there alternatives to the HLM software that NCES should consider using ?

is a resounding "yes".

In the first place, this is a "yes" in the general sense. The two-step ordinary and weighted least squares methods deserve, at the very least, some additional study. The nonparametric and semiparametric methods, and the path analysis and latent variable versions of the multilevel models should also be studied in detail. And, perhaps most importantly, software should be developed that studies the deviations from the multilevel model, preferably in a graphical and interactive way.

Secondly, it is a "yes" in the narrow sense. As far as algorithms for maximum likelihood estimation are concerned, the alternatives are clear. We can choose between the scoring method in Longford's VARCL, the iterative generalized least squares (IGLS) methods in Goldstein's ML/2 and ML/3, and the EM algorithm in HLM and Mason's GENMOD. It is also obvious that there is no uniformly best method, and that none of the three may provide the final answer. In a recent paper Mak [43] proposes a method that does not require much more computation than a single EM iteration, and that basically finishes after one or two iterations.

If we compare advantages and disadvantages, then EM has global convergence from any starting point to a solution which is always feasible (no negative variances). This advantage, however, is also its undoing in other situations. Global convergence means small steps, and thus slow convergence. If there is convergence to a boundary point, EM slows down to a crawl, and it will not get there in our lifetime.

Technically, EM becomes sublinear in such circumstances. The user will have stopped long before this, at a point which looks stationary because nothing is really changing any more. Since EM typically does not give information about the quadratic component of the likelihood function in the region in which it meanders, there is very little information available that can be used to diagnose this situation. Scoring is often said to have locally quadratic convergence. But this is true only if the model is true, which it is not, and if convergence is not to a boundary point or a point where the information matrix is singular. In examples that are ill-conditioned, VARCL also slows down and becomes linear or worse. Both VARCL and IGLS, however, give better indications that something is wrong. Variances become negative, inverses explode, and so on.

From the results of [39] we conclude that VARCL is more difficult to use than HLM, but it gets one to the same solution faster if the model is well-conditioned. If the model is way off, then VARCL has better ways of showing this. More or less the same thing is true for ML/3, but ML/3 is really an interactive software package with a much more general range than HLM. Within ML/3 we can study residuals, compute summary statistics, make plots, and so on. The learning curve is much steeper, but this is unavoidable. Even steeper learning curves result, if the user decides to write multilevel software in Xlisp-Stat or S-Plus, the interactive statistical environments that are rapidly becoming more popular. This gives the maximum amount of user control, but requires also the maximum amount of prior knowledge.

To put it somewhat differently, the HLM program assumes from the start that the basic model is correct, and the number of variations and tests within the basic model that can be tried out is consequently quite limited. Clearly, the developers of HLM have a different class of "users" in mind. Since NCES also has to deal with more sophisticated users, who want to explore their data, experiment with models, and investigate the residuals, we think ML/3 should be available as well. We see the simple order, in terms of *precookedness,*

$$HLM \geq VARCL \geq ML/3 \geq XLISP.$$

This ordering implies that the programs cater to different groups of users.

8. History

In this section we collect some remarks about the history of multilevel models. This completes and improves the remarks on history given in

[19]. We shall try to give the key papers in various areas, and to show what their relationships are. In particular, we concentrate on review papers and textbooks. The purpose of this section is to show that developments similar to the ones in educational statistics are going on, or have been going on, elsewhere as well. Tools developed in one area can often be used in other areas as well. We think one of the useful functions of statistics as an academic discipline is to coordinate and document data analysis developments that are going on in different disciplines.

8.1. Variance Components. Variance component analysis (and mixed model analysis) has a long and complicated history, which is discussed in considerable detail in the book by Searle [55]. The first use of the technique was in astronomy by Airy around 1860. But, of course, the seminal work was by Fisher in his basic 1918 quantitative genetics paper, and in his 1925 book. The distinction between fixed effects and random effects, and the birth of the mixed model, can be dated to the work of Eisenhart around 1945. Between 1950 and 1970 the field was dominated by the Henderson methods for estimating the variance components, and around 1970 the computational revolution made it possible to compute maximum likelihood estimates (Hartley, Rao, Hemmerle, Harville, Thomson, Searle). Since 1970 there has been a lot of emphasis on computation, for which we refer to the excellent review paper by Engel [24], and some progress towards a deeper understanding of what we mean by an "analysis of variance". Two interesting papers on this last topic are by Speed [57] and Samuels, Casella, and McCabe [51].

8.2. Random Coefficients. We have seen that random coefficient models were proposed in econometrics in the Cowles Commission days by Wald in 1947 and by Rubin in 1950. The computational revolution made these models also practically relevant, and in the early seventies there were review papers of Rosenberg [50], Spjøtvoll [62], and a monograph by Swamy [60]. A bibliography has been published by Johnson [36], [37]. Recently there have been some attempts to make random coefficient models semiparametric, in the sense that the distribution of the random effects is not assumed to be normal, but is estimated from the data. In the linear case, see [5], for the nonlinear case, see [18].

8.3. Variable Coefficients. This is a very general class of models which can be described by

$$\underline{y}_i = x_i'\beta(z_i,\theta) + \underline{\epsilon}_i, \tag{43}$$

Each individual has her own vector of regression coefficients, which depends on a number of parameters, possibly in a nonlinear way. In this generality the model depends heavily on computational tools such as smoothing. It has been discussed recently by Hastie and Tibshirani [31], and related to the generalized additive models they discuss in their book [30]. Observe that the coefficients in these models are fixed.

8.4. Changing Coefficients. Consider the following random coefficient model, where the index t replaces i and now stands for time-points.

$$\underline{y}_t = x_t'\underline{\beta}_t + \underline{\epsilon}_t, \tag{44}$$

$$\underline{\beta}_t = M\underline{\beta}_{t-1} + \underline{\eta}_t. \tag{45}$$

Thus regression coefficients satisfy an autoregressive path model. There is a lot of recent interest in this model. The literature until 1984 is reviewed by Chow [13]. There is a close relationship with the Kalman filter of control system theory fame.

8.5. Panel Data. In economics, at least micro-economics, panel data, which follow a number of individuals in time, have received a great deal of attention. We refer to the review paper by Chamberlain [11], and the books by Hsiao [33] and Dielman [22]. The models are usually variable coefficient regression models, sometimes with random coefficients. In many cases they are fairly straightforward mixed models or variance components models [65].

8.6. Growth Curves and Repeated Measurements. Growth curve models have been studied in biometry since Wishart. The key paper here is Pothoff and Roy [46]. They introduce the model $\underline{Y} = X\Gamma Z + \underline{E}$, which can actually be written as a balanced version of the two-level model (19)-(20), without the random component at the second level. Rao [48] linked growth curves with random coefficient modeling. The MANOVA approach to growth curve modeling, and related modeling of repeated measurements, is discussed in the Handbook chapters by Geisser [26] and Timm [61]. The relation with multilevel models is discussed in detail in [58] and [34].

8.7. Bayesian Linear Models and Empirical Bayes Estimation. There is a strong formal relationship between multilevel modeling and the Bayesian analysis of the linear model discussed extensively by Lindley, Smith, Leamer, Zellner and others [41], [56]. We call the relationship "formal" because there is nothing inherently Bayesian about

assuming coefficients to be random. The models can be interpreted equally well as frequentist mixture models.

The use of shrinkage estimators in linear models can also be motivated from mean-square-error considerations, using the basic James-Stein theory. Classical papers by Efron and Morris [23], [45] explain the data analysis aspects of shrinkage estimation. A recent National Research Council report discusses the notion of "borrowing strength" in considerable detail [15]. The report concentrates on meta-analysis as the main area of application, but the methodological discussion is quite general.

8.8. Moderator Variables. The concept of a moderator variable is not easily defined. There is a thoughtful review in [3]. Velicer [64] discusses the concept in terms of different-regressions-in-different-groups, and in an early paper Saunders [52] explicitly takes the point of view that regression coefficients in an equation are themselves dependent variables in a second set of equations.

8.9. Slopes as Outcomes. "Slopes-as-outcomes" analysis was proposed in the late seventies by Burstein and his co-workers as an alternative to the variance decomposition techniques of Cronbach. A nice historical review of the approach is in [9]. The technique is two-step OLS, but it was quite unclear what the precise statistical model behind the computations was. In a sense, the random coefficients models are one attempt to make slopes-as-outcomes rigorous.

REFERENCES

1. L. S. Aiken and S. G. West, *Multiple regression: Testing and interpreting interaction*, Sage Publications, Newbury Park, CA, 1991.
2. Y. Aragon, *Random Variance Linear Models: Estimation*, Computational Statistics Quarterly **1** (1984), 295–309.
3. R. M. Baron and D. A. Kenny, *The Moderator-Mediator Variable Distinction in Social Psychological Research: Conceptual, Strategic, and Statistical Considerations*, Journal of Personality and Social Psychology **51** (1986), 1173–1182.
4. R. J. Beckman, C. J. Nachtsheim, and R. D. Cook, *Diagnostics for Mixed-Model Analysis of Variance*, Technometrics **29** (1987), 413–426.
5. R. Beran and P. Hall, *Estimating Coefficient Distributions in Random Coefficient Regressions*, The Annals of Statistics **20** (1992), 1970–1984.
6. L. H. Boyd and G. R. Iversen, *Contextual analysis: Concepts and statistical techniques.*, Wadsworth, Belmont, CA, 1979.
7. A. S. Bryk and S. Raudenbush, *Hierachical linear models for social and behavioral research: Applications and data analysis methods*, Sage Publications, Newbury Park, CA, 1991.

8. A. S. Bryk, S. W. Raudenbush, M. Seltzer, and R. T. Congdon, *An introduction to HLM: Computer program and user's guide*, University of Chicago, 1988.
9. L. Burstein, K.-S. Kim, and G. Delandshere, *Multilevel investigation of systematically varying slopes: Issues, alternatives, and consequences*, Multilevel Analysis of Educational Data (New York, NY) (R. D. Bock, ed.), Academic Press, 1989.
10. L. Burstein, R. L. Linn, and F. J. Capell, *Analyzing Multilevel Data in the Presence of Heterogeneous Within-Class Regressions*, Journal of Educational Statistics **3** (1978), 347–383.
11. G. Chamberlain, *Panel Data*, HandBook of Econometrics, Volume 2 (Amsterdam, Netherlands) (Z. Griliches and M. D. Intriligator, eds.), North Holland Publishing Company, Amsterdam, Netherlands, 1984.
12. G. Chamberlain and E. E. Leamer, *Matrix Weighted Averages and Posterior Bounds*, Journal of the Royal Statistical Society B **38** (1976), 73–84.
13. G. C. Chow, *Random and Changing Coefficient Models*, HandBook of Econometrics, Volume 2 (Amsterdam, Netherlands) (Z. Griliches and M. D. Intriligator, eds.), North Holland Publishing Company, Amsterdam, Netherlands, 1984.
14. R. Christensen, L. M. Pearson, and W. Johnson, *Case-Deletion Diagonostics for Mixed Models*, Technometrics **34** (1992), 38–45.
15. National Research Council, *Combining Information. Statistical Issues and Opportunities for Research*, National Academy Press, Washington, DC, 1992.
16. D. R. Cox, *Interaction*, International Statistical Review **52** (1984), 1–31.
17. N. Cressie and S. N. Lahiri, *The Asymptotic Distribution of REML Estimators*, Journal of Multivariate Analysis **45** (1993), 217–233.
18. M. Davidian and A. R. Gallant, *NLMIX: A Program for Maximum Likelihood Estimation of the Nonlinear Mixed Effects Model with a Smooth Random Effects Density*, Department of Statistics, North Carolina State University, Raleigh, NC, 1992.
19. J. de Leeuw and I. G. G. Kreft, *Random Coefficient Models for Multilevel Analysis*, Journal of Educational Statistics **11** (1986), 57–86.
20. ———, *Multilevel Path Analysis*, Unpublished notes, UCLA Statistics, Los Angeles, 1993.
21. J. de Leeuw and G. Liu, *Augmentation Algorithms for Mixed Model Analysis*, Preprint 115, UCLA Statistics, Los Angeles, CA, 1993.
22. T. E. Dielman, *Pooled cross-sectional and time series data analysis*, Marcel Dekker, New York, NY, 1992.
23. B. Efron and C. N. Morris, *Data Analysis using Stein's Estimator amd its Generalizations*, Journal of the American Statistical Association **74** (1975), 311–319.
24. B. Engel, *The Analysis of Unbalanced Linear Models with Variance Components*, Statistica Neerlandica **44** (1990), 195–219.
25. P. R. Fisk, *Models of the Second Kind in Regression Analysis*, Journal of the Royal Statistical Society B **29** (1967), 266–281.
26. S. Geisser, *Growth Curve Analysis*, HandBook of Statistics, Volume 1 (Amsterdam, Netherlands) (P. R. Krishnaiah, ed.), North Holland Publishing Company, Amsterdam, Netherlands, 1980.

27. A. S. Goldberger, *Best Linear Unbiased Prediction in the Generalized Linear Regression Model*, Journal of the American Statistical Association **57** (1962), 369–375.
28. E. A. Hanushek, *Efficient Estimates for Regressing Regression Coefficients*, American Statistician **28** (1974), 66–67.
29. D. A. Harville, *Maximum Likelihood Approaches to Variance Component Estimation and Related Problems*, Journal of the American Statistical Association **72** (1977), 320–340.
30. T. Hastie and R. Tibshirani, *Generalized additive models*, Chapman and Hall, London, GB, 1990.
31. ———, *Varying Coefficient Models (with discussion)*, Journal of the Royal Statistical Society B **55** (1993), 757–796.
32. J. Hemelrijk, *Underlining Random Variables*, Statistica Neerlandica **20** (1966), 1–7.
33. C. Hsiao, *Analysis of panel data*, Cambridge University Press, Cambridge, GB, 1986.
34. R. Jennrich and M. Schluchter, *Unbalanced repeated measures models with structured covariance matrices*, Biometrics **42** (1986), 805–820.
35. S. Johansen, *Asymptotic Inference in Random Coefficient Regression Models*, Scandinavian Journal of Statistics **9** (1982), 201–207.
36. L. W. Johnson, *Stochastic Parameter Regression: an Annotated Bibliography*, International Statistical Review **45** (1977), 257–272.
37. ———, *Stochastic Parameter Regression: an Additional Annotated Bibliography*, International Statistical Review **48** (1980), 95–102.
38. L. R. Klein, *A textbook of econometrics*, Row, Peterson and Company, Evanston, IL, 1953.
39. I. G. G. Kreft, J. de Leeuw, and K.-S. Kim, *Comparing Four Different Statistical Packages for Hierarchical Linear Regression: GENMOD, HLM, ML2, VARCL.*, Preprint 50, UCLA Statistics, Los Angeles, CA, 1990.
40. N. Lange and L. Ryan, *Assessing Normality in Random Effects Models*, The Annals of Statistics **17** (1989), 624–642.
41. D. V. Lindley and A. F. M. Smith, *Bayes Estimates for the Linear Model*, Journal of the Royal Statistical Society **B34** (1972), 1–41.
42. N. T. Longford, *VARCL. Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood)*, Educational Testing Service, Princeton, NJ, 1990.
43. T. K. Mak, *Solving Non-linear Estimation Equations*, Journal of the Royal Statistical Society **55** (1993), 945–956.
44. J. J. Miller, *Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance*, The Annals of Statistics **5** (1977), 746–762.
45. C. N. Morris, *Parametric Empirical Bayes Inference: Theory and Applications*, Journal of the American Statistical Association **78** (1983), 47–65.
46. R. F. Pothoff and S. N. Roy, *A Generalized Multivariate Analysis of Variance Model Useful Especially for Growth Curve Problems*, Biometrika **51** (1964), 313–326.

47. J. Rabash, R. Prosser, and H. Goldstein, *ML2. Software for Two-Level Analysis. User's Guide.*, Institute of Education, University of London, London, GB, 1989.
48. C. R. Rao, *The Theory of Least Squares when Parameters are Stochastic and its Application to the Analysis of Growth Curves*, Biometrika **52** (1965), 447–458.
49. G. K. Robinson, *That BLUP is a Good Thing: the Estimation of Random Effects (with discussion)*, Statistical Science **6** (1991), 15–51.
50. B. Rosenberg, *A Survey of Stochastic Parameter Regression*, Annals of Economic and Social Measurement **2** (1973), 381–397.
51. M. L. Samuels, G. Casella, and G. P. McCabe, *Interpreting Blocks and Random Factors (with discussion)*, Journal of the American Statistical Association **86** (1991), 798–821.
52. D. R. Sauders, *Moderator Variables in Prediction*, Educational and Psychological Measurement **16** (1956), 209–222.
53. S. R. Searle, *Linear models*, Wiley, New York, NY, 1971.
54. ———, *Topics in Variance Component Estimation*, Biometrics **27** (1971), 1–76.
55. S. R. Searle, G. Casella, and C. E. McCulloch, *Variance components*, Wiley, New York, NY, 1992.
56. A. F. M. Smith, *A General Bayesian Linear Model*, Journal of the Royal Statistical Society B **35** (1973), 67–75.
57. T. P. Speed, *What is an Analysis of Variance (with discussion)*, The Annals of Statistics **15** (1987), 885–941.
58. J. L. F. Strenio, H. I. Weisberg, and A. S. Bryk, *Empirical Bayes Estimation of Individual Growth Curve Parameters and their Relationship to Covariates*, Biometrics **39** (1983), 71–86.
59. W. H. Swallow and J. F. Monahan, *Monte Carlo Comparison of ANOVA, MIVQUE, REML, and ML Estimators of Variance Components*, Technometrics **26** (1984), 47–57.
60. P. A. V. B. Swamy, *Statistical inference in a random coefficient model*, Springer, New York, NY, 1971.
61. N. H. Timm, *Multivariate Analysis of Variance of Repeated Measurements*, HandBook of Statistics, Volume 1 (Amsterdam, Netherlands) (P. R. Krishnaiah, ed.), North Holland Publishing Company, Amsterdam, Netherlands, 1980.
62. E. Spjøtvoll, *Random Coefficients Regression Models. A Review*, Mathematische Operationsforschung und Statistik **8** (1977), 69–93.
63. R. van der Leeden, K. Vrijburg, and J. de Leeuw, *A Review of two Different Approaches for the Analysis of Growth Data Using Longitudinal Mixed Linear Models: Comparing Hierarchical Linear Regression (ML/3, HLM) and Repeated Measures Design with Structured Covariance Matrices (BMDP-5V)*, Preprint 98, UCLA Statistics, Los Angeles, 1991.
64. W. F. Velicer, *The Moderator Variable Viewed as Heterogeneous Regression*, Journal of Applied Psychology **56** (1972), 266–269.
65. T. J. Wansbeek, *Quantitative effects in panel data modelling*, Ph.D. thesis, University of Leiden, 1980.
66. M.B. Wilk and O. Kempthorne, *Fixed, Mixed, and Random Models*, Journal of the American Statistical Association **50** (1955), 1144–1167.

Departments of Psychology and Mathematics UCLA, 405 Hilgard Avenue, Los Angeles, CA 90024-1555
E-mail address: deleeuw@stat.ucla.edu

School of Education CSLA, 5151 State University Drive, Los Angeles, CA 90032-8143
E-mail address: kreft@stat.ucla.edu