DOCUMENT RESUME

ED 371 033 TM 021 737

AUTHOR Kreft, Ita G. G.; Yoon, Bokhee
TITLE Are Multilevel Techniques Necessary? An Attempt at Demystification.
PUB DATE Apr 94
NOTE 19p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 4-8, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Correlation; Data Analysis; *Educational Research; *Effective Schools Research; Elementary Secondary Education; Error of Measurement; *Estimation (Mathematics); Instructional Effectiveness; Models; Monte Carlo Methods; Prediction; *Regression (Statistics); Research Problems; Statistical Studies
IDENTIFIERS *Multilevel Analysis; *Random Line Models

ABSTRACT

The merits of the multilevel model for educational research and its uses for school effectiveness research are considered. The main goal of the paper is to establish what intelligent applications of multilevel models can do, helping researchers decide what they must do to make a rational choice between models. Multilevel models are random line models. The data is assumed to be hierarchically nested, with the lower observations nested within the higher levels, resulting in intraclass correlation. Random line models are random coefficient, fixed variable models. The parameters of the models and some alternative estimation procedures are explored. Monte Carlo studies illustrate the use of these models. It is concluded that the random line model is technically an improvement over the traditional multiple regression model because it calculates the correct standard errors. In addition, researchers interested in prediction for separate schools will find the random line model technically superior because the procedure improves the estimation of the parameters for the separate schools. Random line models can be useful, but will often lead to the same conclusions as classical regression models. (Contains 38 references.) (SLD)

ARE MULTILEVEL TECHNIQUES NECESSARY ? AN ATTEMPT AT DEMYSTIFICATION

ITA G.G. KREFT AND BOKHEE YOON

April 1, 1994

> We see a bunch of facts, as our point of view changes, so do our observations. And, after some computer-work, linear regression or what-not, we conclude "That we have solved the mystery as convincingly as a Sherlock Holmes"
> *Guillaume Wunsch,* [38]

1. INTRODUCTION

1.1. School Effectiveness Research and the Merits of a Multilevel Model. "In school effectiveness research we observe that many studies have failed to obtain reasonable evidence of a nonzero school, classroom or teacher effect. This has been attributed to methodological errors leading to false conclusions." (Rachman-Moore and Wolfe, [31], p.277). The newly developed multilevel models and techniques contain the promise of less methodological errors. But models are never true or false, and so are techniques never true of false. It is the use of the model and the techniques that decides its merits.

This paper is a critical review of the merits of the multilevel model and its uses for school effectiveness research. We wonder if promises made like *enriched discussion* and *advancement of research in school effects* have been and can be fulfilled by using multilevel models. Have multilevel modelers demonstrated where researchers using traditional models have erred in the past? Do they give some understanding of that what we did not know before? Since multilevel models are here to stay, its is time that the users of such models become aware that no model by itself can enhance advancement of the field. "This type of program appeals to many social scientists, who are very unsure about the value of their prior knowledge. They prefer to delegate

Paper to be presented at the Annual Conference of the American Educational Research Association April 4-11, 1994, New Orleans.

Special thanks to Rien van der Leeden and Kyung-Sung Kim for their assistance with the Monte Carlo results, and to Jan de Leeuw for TEXpertise.

DRAFT VERSION: DO NOT CITE OR QUOTE.

decisions to the computer, and they expect techniques to generate knowledge." is a citation from De Leeuw [12] regarding LISREL models, and that can be used here again. Since multilevel models are not the panacea for all the methodological and conceptual problems that has plagued the field, we want to explore what the model can and cannot do and what researchers can expect of this type of model. After all the shrunken estimates resulting from this model show less instead of more school effects. It is not always clear in which situations the model is able to reveal new knowledge.

Multilevel model are intuitively very appealing. A practical appeal is the concept of *borrowing of strength*, which is relevant when there is sparse data (like data on minorities applying for a special law school, see Rubin, [32]). A statistical appeal is that it promises to improve the estimation of individual effects. And last but not least the model invites to fit cross level interactions, which links the model to theories in education (see Cronbach and Webb, [11] and Cronbach and Snow, [10]). In Cronbach's theory it is argued that teachers styles differ and that some styles are more effective for one group of students (low achievers for instance) than for another. As a result, a teacher effect could and should be measured as an interactive effect between the qualities of the teacher and the qualities of the student. A more recent example is Aitkin and Zuzovsky [2], who declare the 1990's research framework for school effectiveness research as *interactive*. Interactive is defined as differential effectiveness for specified subpopulations affected by particular treatment under special conditions. The appropriate statistical model is nonadditive with conditional dependence of achievement on a combination of variables interacting with each other. Bryk and Raudenbush [6] agree when they see as one purposes of the model (l.c. p.5) "The formulation and testing of the hypotheses about cross-level interactions".

1.2. Goal and Purpose of this Paper. Our goal is to establish the merits of multilevel models for school effectiveness research. These merits can be established in two respects: a technological and a conceptual one. A technological merit is when parameters are measured more precise and non biased. A conceptual merit is that the models provide better predictions, and advance the knowledge regarding school effectiveness. De Leeuw, in the introduction of Bryk and Raudenbush [6], sees the main gain of the model as conceptual.

1.3. A technological issue: The choice between a general and a more restricted model. An obvious merit of the multilevel model is that it is based on assumptions that are more realistic. In the social sciences we have the situation that relations and interactions between people are not well understood or defined. Few theories exist that elaborate the way the data are generated. In such uncertain situations we have the choice between two options: the use of the most general model and let the data decide, or the use of a more restricted model. This is the choice

between the more general multilevel model, or the restricted OLS regression model. While it is true that a more general model is preferred to a more restricted model in uncertain situations, large data sets are needed to prevent instability of the solutions in the more complicated multilevel model. An interesting question is, if the more general model is necessary to improve estimation, and to enhance theories of school effectiveness.

1.4. A conceptual issue: Intelligent applications of multilevel models. "If the intelligent application of multilevel models promises a deeper understanding of the processes of schooling and the determinants of achievement" (Goldstein [17], p.90), we need to know first what intelligent applications are before we can come to a better understanding of the world based on multilevel analyses. Besides, the merit of a model for advancing social theory relies mainly on our knowledge of the data and the field we study. If that knowledge warrants the use of multilevel analysis it still cannot prevent us of continuing to make non-replicable inferences about school effectiveness. As Longford [24] argues, a substantial element of artifact is contained in the definition of the *school effect* in terms of outcomes of certain statistical procedures. This is the result of both, the way the data are collected by non random sampling and the observational character of this type of studies. The multilevel model is not discussed in relation to its potential to theory building, because we agree with Longford that given the procedures of data collection, such claims are questionable. The model will be discussed in relation to its power, its efficiency, and how it alters and enriches the discussion in school effectiveness research.

1.5. Situations where Multilevel models are clearly needed. The merit of the multilevel model has been established in situations with high intra class correlation, such as identical twin studies, where an intra class correlation of 0.90 is observed. A natural application of the multilevel model is in the measurement of progress, such as repeated measurements on students and changes of schools over time. For instance in studying change of success of schools, multilevel models produce more reliable estimates than fixed regression models, as a result of shrinkage, and borrowing of strengths with sparse data. Shrinkage is employed in EB/ML estimation procedures to correct for unreliable estimates. The partitioning of the variance in a within and between part is most relevant in these examples, since the between part is substantial and cannot be ignored without loss of efficiency. Shrinkage makes a case for better generalizability in some clearly defined cases (Goldstein, [17], Bryk and Raudenbush, [6]). In Bock [5] several examples show the merits of the model for prediction purposes, where sparse data are present (Rubin, [32]) and the number of observations per school or department is too small to make good predictions. How much better the shrunken estimates do in comparison to the pooled Least Squares (LS) estimates, is not clear. Rubin [32] finds that the Bayesian estimates are better predictors, but

the difference is small. An entirely different matter is if the observed difference is large enough to provide us with more opportunities for better conceptualization.

Our main goal in this paper is to establish what *intelligent applications* of multilevel models are. In that way we can help researchers to establish what necessary prior information they need *to make a rational choice* between models, as is suggested by De Leeuw in his introduction to Bryk and Raudenbush ([6], p XVI)): "Social statisticians will be able to do more extensive modeling, and they will be able to choose from a much larger class of models. If they are able to build up the necessary prior information to make a rational choice from the model class, then they can expect more power and precision. It is a good idea to keep this in the back of your mind as you use this book to explore this new exciting class of techniques."

In the next paragraphs the multilevel model is defined and a suitable name is found in order to distinguish the multilevel model we are talking about from other multilevel models that are around.

2. Definition of Multilevel Models and Multilevel Techniques

2.1. Multilevel models as *random line* models. The term *Multilevel Model* needs a more specific definition, because all models handling multilevel data are multilevel models. In multilevel models the data is assumed to be hierarchically nested, where the lower observations are nested within higher levels, resulting in intra class correlation. In this paper we report simulation studies where the data are generated by a random coefficient process. Software packages for the analysis of this type of data are HLM (Bryk, Raudenbush, Seltzer, Congdon, [7]), ML3 (Prosser, Rasbash, and Goldstein, [30]), VARCL (Longford, [25]). The models underlying the software are based on the notions of separate analyses for each group, where the coefficients obtained from these analyses are used as dependent variables in an aggregated analysis together with group level variables. This is based on the *slopes-as-outcomes* approach as developed in the seventies (e.g. Burstein, Linn and Capell, [8]).

Since the name *slopes as outcomes* is reserved for the separate fixed regression analysis described above, we propose to use the name *random line* models for the type of data analysis we discuss here. This distinguish it from multilevel structural equations models such as BIRAM (see McDonald, [28]), and a version of LISCOMP (see Muthen, [29]). Latter models are random variable, fixed coefficient models, while *random line* models are random coefficient, fixed variable models. The name distinguishes it also from a software package for the analysis of this type of data, HLM, an acronym for Hierarchical Linear Model (Bryk, Raudenbush, Seltzer, Congdon, [7]). Instead of *slopes-as-outcome* model, multilevel model or Hierarchical Linear Model, the term *random line* model is used throughout this paper. The name is adopted from McDonald [27], who uses the name for the same purpose as we do, to distinguish

the two different multilevel models, mentioned here, from each other.

2.2. Parameters of the *Random Line* model. For a better understanding we introduce the parameters of *random line* model. In the following notation, where underlining indicates a random variable, X is the predictor and $\underline{Y}$ the dependent variable. Index i is used for individuals, index j is used for contexts.

$$\underline{Y}_{ij} = \underline{a}_j + \underline{b}_j X_{ij} + \underline{e}_{ij} \tag{1}$$

$$\underline{a}_j = \gamma_{00} + \gamma_{01} Z_j + \underline{u}_{0j} \tag{2}$$

$$\underline{b}_j = \gamma_{10} + \gamma_{11} Z_j + \underline{u}_{1j} \tag{3}$$

The macro-level errors (disturbances) $\underline{u}_{0j}$ and $\underline{u}_{1j}$, in (2) and (3) respectively, indicate that both the intercept $\underline{a}_j$ and $\underline{b}_j$ vary over contexts. The grand mean effect in (2) is γ_{00}, while $\underline{u}_{0j}$ (the macro-error term) measures the deviation of each context from this overall or grand mean. The same is true in (3) where the grand slope estimate across all contexts is γ_{10}, while $\underline{u}_{1j}$ represents the deviation of the slope within each context from the overall slope. For the gammas the subscript is defined as follows: the first index is the number of the variable at the micro level, the second represents the number of the variable at the macro level. Thus γ_{st} is the effect of the macro level t on the regression coefficient of micro variable s. Zero signifies the intercept, i.e. the variable with all values equal to +1, either at the micro level or at the macro level. For instance γ_{00} is the effect of the macro level intercept on the micro level coefficient of the intercept. Note that (2) and (3) display the model coefficients $\underline{a}_j$ and $\underline{b}_j$ as a function of two components: a fixed component γ_{00} and γ_{10} respectively, and a random component $\underline{u}_{0j}$ and $\underline{u}_{1j}$ respectively, where $\underline{u}_{0j}$ has variance τ_{00} , $\underline{u}_{1j}$ has variance τ_{11}, while $\underline{u}_{0j}$ and $\underline{a}_{1j}$ have covariance τ_{01}. Macro-level variables Z can be introduced into the equations for the intercept and/or slope (i.e. (2) and (3) above). In (2) the intercept $\underline{a}_j$ of each context is now shown to be a function of both the group level variable Z_j and random fluctuation $\underline{u}_{0j}$. The same happens in (3), where the slope is a function of the same group level variable.

The variances of the macroerrors $\underline{u}_{0j}$ and $\underline{u}_{1j}$ and their covariance are parameters of the model, and are given in the matrix T. The terms in T are referred to as variance components of the model. For the taus the subscripts all refer to macro level variables. This means that τ_{st} is the covariance between random regression coefficients s and t. Zero refers again to the random intercept.

$$T = \begin{array}{c} \\ \delta_{0j} \\ \delta_{1j} \end{array} \begin{array}{c} \begin{array}{cc} \delta_{0j} & \delta_{1j} \end{array} \\ \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix} \end{array} \tag{4}$$

3. Alternative Estimation Procedures

Procedures used to estimate parameters in regression models with group data are either unweighted least squares, such as OLS, or weighted least squares, such as Generalized Least Squares (GLS) or in EB/ML , where at each iteration the weights are adjusted and changed.

OLS estimates are used in multilevel software as the starting values for the iterative process. OLS estimates ignore the nested structure of the data. WLS or GLS is the outcome after one single iteration, using estimates for the error structure for all levels based on the OLS estimates. The equations for OLS and WLS are given below.

The OLS and GLS estimators are all of the same form. The covariance of the estimates is

$$(\sum_{j=1}^{m} W_j' S_j W_j)^{-1} (\sum_{j=1}^{m} W_j' S_j W_j S_j W_j) (\sum_{j=1}^{m} W_j' S_j W_j)^{-1} \tag{5}$$

Where

$$W_j = T + \sigma^2 (X_j' X_j)^{-1} \tag{6}$$

and S_j is different for OLS and WLS. For OLS

$$S_j = X_j' X_j, \tag{7}$$

and for WLS

$$S_j = W_j^{-1}. \tag{8}$$

. The weights W_j in Weighted Least Squares can either lead to the procedure of GLS (for instance by employing Swamy weights, see De Leeuw and Kreft, [13]) or are weights that change in an iterative process as in the EB/ML procedure or in Iterative Generalized Least Squares (IGLS). The two methods (iterative and non iterative) are expected to give different results (based on asymptotic theory), but the extent and the direction of the difference is not known in many practical situations. For more details on estimation methods see Chapter 2 in Goldstein [16] and Chapter 3 in Bryk and Raudenbush [6].

In 1986 when we published our first article regarding multilevel modeling (De Leeuw and Kreft, [13]) our conclusion was, that several different procedures (Table 2 and 3, p. 77) gave very similar parameter estimates. The procedures employed are: OLS two step and one step, weighted Least Squares (GLS, with Swamy weights) and WLS with ML (EB/ML). The reported parameters are the three coefficients for the micro variables IQ, sex, SES, one for the macro variable: mean IQ, and their three cross level interactions. Our main conclusion was than that (l.c., p.79) "choice of

estimation method does not seem to have much influence on the size of the regression coefficients." The standard errors show in the same way over all procedures that the important estimates are estimated with roughly the same precision. We concluded that the Rao- Swamy-Johansen weighted least-squares method seemed an excellent method to estimate the unknown parameters of the model, at least when the estimates of the variances and covariance components (the second level disturbances) are not too negative. These early conclusions are reexamined in this paper.

4. Research Questions

Random line models have been around for a decade now. Of course the advice that they should be applied whenever data are hierarchically nested is too general. We need to know when *random line* models are making a difference and for what purpose: prediction, policy decisions or theory development and/or hypothesis testing? For prediction and policy decisions the parameters of importance are the gammas and their respective standard errors. Variance components are interesting from a theoretical point of view.

In this paper three questions are addressed:

(1) Are multilevel models enhancing theory?
(2) Are the OLS estimates in regression models always inappropriate, biased and imprecise?
(3) What is the strength of multilevel models: prediction or description and exploration of trends in the data?

4.1. Do these new models enhance our understanding of the world? Researchers seem to think of multilevel research either as a tool for hypothesis testing, or as a descriptive/explorative tool. On a scale ranging from very careful formulation of the results of a *random line* model to a very prudent one, we could say that Longford is on the low end (the careful one). Ranking, in ascending order: Longford [24] concludes that this model provides a suitable framework for data description in a variety of contexts, but he believes that the results cannot be used for explanatory purposes. Goldstein [16] sees the model as an explorative tool for theory development as is clearly stated in the introduction to his book: (p. 8):"This book is ... for the explicit purpose of exploring and explaining relationships and variations within and between the levels of the system." He continues: (p.9): "Their (the models) use in this book in no way constitutes a recommendation to particular problems." The book by Bryk and Raudenbush [6] states as research purposes of the model (p.5): the improved estimation, the formulation and testing of hypotheses and the partitioning of variance. The inappropriateness of causal statements based on observational data is not discussed here. Interested readers are referred to Draper [14].

The value of these models are seen as, at least, stimulating theories, by altering the way researchers think about analyzing their data (e.g. Aitkin and Longford, [1], Burstein, Linn and Capell, [8], Tate and Wongbundit, [35], and Aitkin and Zuzovsky, [2]). The focus in this paper is on interaction theory. Are cross level interaction effects, as hypothesized by Cronbach et al. ([11] and [10]), detectable with *random line* models? Are the models powerful in that respect? And if so, are the new techniques better equipped for detecting such aptitude $\times$ treatment effects than the traditional methods? Policy makers, researchers and administrators would like to know what factors are associated with success or failure of schools. Can *random line* models give these answers, or are these models merely research tools for exploring the nature of school effectiveness? This leads to the following question:

4.2. How good are the estimates ? Do procedures exist that are equally good or better compared to EB/ML estimates for answering questions as asked by policy makers and administrators? Where good is defined as producing parameter estimates as close as possible, on average, to the true values, and with a small as possible variation around that value. We express some doubts about the claims of superior versus inferior estimation procedure, based on the following studies.

Tate and Wongbundit [35] show that the separate equations model (also called the two steps model) is superior for statistical inferences when data are generated with random coefficients processes. Practical disadvantages are seen as: costly, cumbersome and lacking an appropriate way to determine standard errors for effects in more complex models. They call for "a single unified analysis that still allows for random within-class coefficients" (Tate and Wongbundit, [35], p. 118). De Leeuw and Kreft [13] show that for their data OLS produces unbiased estimates (when standard errors are weighted). The choice of estimation method does not seem to have much influence on the size of the regression coefficients nor on the weighted standard errors. The really important regression coefficients are estimated with roughly equal precision by all techniques. The small regression coefficients are however estimated more precisely by maximum likelihood and weighted least squares compared to OLS (De Leeuw and Kreft, [13], p. 79). Former conclusion is quite satisfactory because most people in educational research use simple least squares, and it is quite likely that they will continue to do so.

4.3. Description or Prediction. The last question deals with the purpose of the analysis: description or prediction? Can OLS estimates in regression models be considered appropriate estimation techniques for prediction purposes? The question what predicts a student's score better: a regression model or a multilevel model depends on for whom the prediction is made: a new student from a new school (both not included in the sample) or for a student included in the sample. Also: are inferences being made for a particular school, or for general purposes ? Can

the outcome score of a new student from a new school be better predicted with the outcome of an OLS procedure, based on the total data set, than from the shrunken estimates of a comparable school in the sample?

Goldstein [17] warns against using posterior means for prediction or ranking of schools, since the estimates are unreliable (large standard errors) and sensitive to the model assumptions. He suggests to use multilevel models as a descriptive tool or *screening* test for exploring the nature of school effectiveness. Rubin [32] concludes that prediction based on sparse data improves using all available data, either by borrowing of strength, or with OLS estimates, based on the pooled data. Prediction of student achievement based on the shrunken estimates still may be a questionable thing to do. A cautionary note can be found in Bryk and Raudenbush ([6], p. 80).

All these questions deal with the same problem: where and when do we profit from the use of *random line* models? As is the case with other analysis models, the researcher has to defend her choice of model, given the data and given the theory. Claims that EB/ML produce the most efficient estimates are true in theory with large datasets, and normal distributions. With non normal data and small datasets this statement cannot be made without further investigation.

5. Overview of Simulation Studies

5.1 The most efficient estimates. How much ? There are several reasons for such a further study. For instance, Kim [19] remarks, that the EB/ML method requires considerable computation time and effort, and therefore it is important to know when the EB/ML procedure leads to superior solutions. Kims results show that it is not necessary to use a very complicated EB/ML method when we have a sufficiently large number of cases (at least 2000 observations), where sufficient means sufficient within group units as well as sufficient number of groups.

Bassiri [4] claims (p.23) "that if we want to analyze a hierarchical structured dataset, procedures involving random variables or random effects would be more accurate than any other method". We dont know how much more accurate, since Bassiris study does not compare estimation methods. Her study is about estimating the power of a random line model when different number of groups and different number of observations within groups are present. And even if there is a difference, we do not know how large the difference is, and if they will lead to the promised advancement in school effectiveness research.

Mason, Wong and Entwistle ([26]. p.76) conclude that "the fixed effects regression model poses no unusual estimation or computation problems, the fixed effects regression model has been used frequently in multilevel models". But in a footnote at the same page they say: "The OLS standard errors are meaningless numbers if

the analyst believes that the micro parameters are not completely determined by the macro variables." Mason et al.'s suggestion is, that we can use OLS in cases were the variation in the intercepts and slopes are determined by the macro variables, and no random variation exists.

5.2. Power, Bias and Efficiency. In an overview of results obtained from Monte Carlo studies power, bias and efficiency are topics discussed, sometimes in relation to estimation procedures such as EB/ML versus GLS (e.g. Kim, [19], Busing, [9], and Van der Leeden and Busing, [36]). The parameters examined in the papers are: the fixed effect (gammas) and the power to detect these under different circumstances (Kim, [19], and Bassiri, [4]), the efficiency of the gammas over 50 replications, and the bias and efficiency of variance components in multilevel models using unrestricted ML as the procedure (Busing, [9]) or restricted ML (Van der Leeden and Busing, [36]).

These authors also studied the effect of three procedures (OLS, GLS and RIGLS) on the estimation of variance components and a cross level interaction. Bassiri studied power of the models as well as the consistency, bias and efficiency of EB/ML estimators for the gammas for micro, macro and interaction terms. Kim and Yoon studied the difference between the results obtained by estimation of the gammas in OLS, GLS and EB/ML. In the final results bias and efficiency are compared between the parameters estimated with GLS or estimated with EB/ML, under different conditions.

OLS, GLS parameters estimates are compared in bias and efficiency with the EB/ML parameter estimates. These are reported in three studies, Busing [9], Van der Leeden and Busing [36], Kim [19].

5.3. Parameters investigated in the reviewed simulation studies. The parameters in *random line* models are the fixed effects or gammas and the random effects or omegas:

- The gamma estimates: The micro level parameters for slope (γ_{00}), intercept (γ_{10}), and cross level interaction between micro and macro (γ_{11}), the macro parameters for intercept (γ_{01}) and slope (γ_{11}) , and their respective standard errors.
- The variance components are: micro variance (σ^2), and macro variances for intercept (τ_{00}) and slope (τ_{11}) and the covariance between the two (τ_{01}).

5.4. Similarities Between the Studies Reviewed. The simulation studies discussed here have in common that all use balanced data, artificially constructed based on the *random line* model and analyzed with software that have OLS estimates as the starting values.

The data are generated based on a *random line* model, followed by a *random line* analysis (obviously the correct model to employ in this situation, which is at the same time the problematic part of simulation studies). The artificiality of the data leaves us in doubt about the results which would be obtained from data not generated with a *random line* model but collected in real life situations.

In school effectiveness research it is mostly unknown how the data are generated. The assumption of exchangeability in multilevel models is questionable, for instance. Gray [18] gives arguments that put some doubts on this concept by showing how correction for background variables of the student body does not equate the schools or make them interchangeable.

Most studies use balanced data. The exception is Busing [9] and Van der Leeden and Busing [36]. They use slightly unbalanced data. Unbalancedness of the data seem to be of minor importance according to the study by Swallow and Monahan [34]. It shows that unless the data is severely unbalanced, the same conclusion regarding the best estimation procedures are reached with balanced as with unbalanced data.

The initial estimates or starting values in all software packages for *random line* models are OLS estimates. In the package ML3 the OLS estimates are used to compute the first set of random parameter estimates, using the part of the algorithm which carries out GLS for the random parameters as described in Goldstein [16]. In the software VARCL the starting values are for the regression parameters the OLS estimates, and for the variances the moment method estimates. All covariance parameters are initially set to zero. Starting values and values obtained after a sufficient number of iterations are part of the output in all of the four existing software packages for *random line* models (for a comparison of the software see Kreft, De Leeuw and Kim, [23], or Kreft, de Leeuw and Van der Leeden, [22]).

6. Results

6.1. Some Preliminary Findings. First we report some early findings that made us more critical about the merits of multilevel models. In Kreft [20] we found no differences in the parameter estimates (and their respective standard errors) in a regression model using OLS compared to the estimates of the same model based on EB/ML estimates. The study was a reanalysis of Webb [37], with 135 students and 35 groups. The single level path analysis employed by Webb gave the same results as our reanalysis with a multilevel model. The intra class correlation was 0.20, too large to ignore. Another study (Kreft and De Leeuw, [21]) shows again surprising results. The ranking of schools based on a multilevel model produces rankings that correlated 0.99 with rankings obtained with a traditional analysis of variance model. The ranking based on (the more appropriate) analysis of covariance model, with correction for background variables (IQ and social economic status) produced quite

different rankings. The effect of shrinkage on the slope parameters in the multilevel model was so large that the control for the background variables was reduced to zero. Both studies (Kreft, [20], Kreft and de Leeuw, [21]) show that the choice of an analysis model is not dictated by the hierarchically nestedness of the data, but need to be made based on knowledge of the situation and the purpose of the analysis. Predictions based on posterior means are at least questionable in certain situations (Goldstein, [17], Fitz Gibbon, [15]).

Experience with the package VARCL (Longford, [25]) provided more information. In VARCL the OLS starting values are part of the output. Comparison of the starting and finished values showed many times no significant differences. We wonder what the conditions are under which EB/ML estimates improve estimation compared to the OLS estimates. It is accepted knowledge that the standard errors in OLS procedures are systematically under estimated when we are dealing with hierarchically nested data. It is interesting to see to what extend OLS estimates of the standard errors in least squares regression are wrong. It is known that the presence of intraclass correlation produces downward biased standard errors, which on their turn produce too high alpha levels (Barcikowski, [3]). There may be a tradeoff, however. For instance: the sigma squared (the micro level error variance) can be overestimated in OLS, as a result of the fact that all error variance is contributed to the micro level and nothing to the macro level, since this level is ignored. As a result the standard errors of the parameter estimates may be closer to the ML/Bayes estimates than is expected. It is known that the OLS single-equation procedure is optimal if T (the matrix containing the second level disturbances) is zero and if all sigma squared (σ_j^2 is the micro level disturbances within each group) are equal. Thus for small T and for approximately equal σ_j^2 (for instance as a result of a small number of observations in groups, like in the Webb data) , the single equation method will give a good approximation.

6.2. The Estimation of Interactions. We are interested in interactions, because it is the main attractiveness of the model from a theoretical point of view. It contains the promise of *new discoveries and new discussions* regarding school effectiveness. For that reason the number of groups and observations is studied in relation to the power of detecting interaction effects. And three different estimation procedures are compared in the estimation of the interaction effects. The studies are Bassiri [4], Kim [19] and Van der Leeden and Busing [36]. Van der Leeden et al. find no differences between three estimates for an interaction effect obtained by the software ML3 (Prosser et al. [30]). The three estimates are the OLS starting values, the GLS value after one iteration, and the value after convergence is reached, the RIGLS or the IGLS estimate.

RESULTS ABOVE NOT YET COMPLETE

Conditions in this study are an intra class correlation of 0.10 and 0.20, which are the most common intra class correlations in educational research. Sample sizes are 25 different combinations of 10 to 65 groups and 5 to 40 observations within groups. The maximum number of total observations is 2600 (65 groups of 40) and the lowest is 50 (10 groups of 5). The data are not completely balanced (see Bryk and Raudenbush, [6], p.228-229). None of the conditions show any significant differences in results between the estimation procedures in the gamma for the interaction term.

From the study of power by Bassiri [4] we know that interaction effects are not easily detected: to reach a power of 0.90 many groups are needed. The best situation starts with 60 groups with 25 observations per group, or less groups and more observations (30 groups with 150 observations for instance) or more groups and less observations (150 groups with 5 observations each, for instance). More groups seem to be better for interactions and second level parameter estimates, while the first level estimates solely depend on the total number of observations. This study seem to suggest that for multilevel purposes more groups pays off by the fact that we need a lower number of total cases. On the other hand it may be many times more costly to collect more groups than more individuals within a group as Snijders and Bosker [33] show. This article attempts to calculate this trade off in relation to power together with a minimization of cost.

Based on Bassiri it seems that for a model powerful enough to detect a cross level interaction the choice is between 4500 observations in 30 groups, 1500 observations within 60 groups or 750 observations within 150 groups. Using less observations show a rapid decline of power for interaction effects. The conditions of Bassiri's study are an intra class correlation of 0.10 and 0.25. Sample sizes range from 10 groups to 150 groups and from 5 observation per group to 150 observations per group.

6.3. OLS, GLS and RIGLS (or EB/ML) Compared. For this comparison we have to rely on two studies, Kim [19] and Van der Leeden and Busing [36]. Kim and Van der Leeden are described earlier. Yoon calculates the precision of the estimation methods OLS, GLS and EB/ML.

The conclusion of Kim [19] over 50 replications is, that GLS estimates for the gammas are equal in bias and efficiency to EB/ML (obtained by using the software HLM, Bryk et al, [7]). The fixed regression parameters in his study are estimates for first level and second level variables and their interactions. A more detailed reanalysis of Kim's data shows that there are no differences in precision between OLS and the other two estimation methods (GLS and EB/ML). The gamma-estimates are unbiased and equally efficient. The observed variance in the gammas over the fifty replications does not improve after the first iteration. GLS has the same variance as EB/ML after convergence is reached. OLS is less efficient, the variance is larger. The efficiency is 90% compared to GLS and EB/ML. This means that more observations

are needed for OLS to reach the same precision.

Van der Leeden and Busing [36] examine the variance components of the *random line* model, using the software ML3 (Prosser et al. [30]). The conclusions regarding the variance components is completely different from the conclusion for the gammas reached in Kim [19]. The change from GLS to RIGLS is substantial, as measured in a repeated-measures MANOVA. All interactions between the estimates and the conditions are significant. Interaction effects in the models are between the two different procedures (GLS versus RIGLS) and the number of groups (J), the number of observations within groups (N_j) and the two different intraclass correlations ($r = 0.10$ and $r = 0.20$).

Van der Leeden and Busing again compared the starting values for the variance components (the GLS estimates) with the unrestricted IGLS version of ML3. The same changes occur between the two estimation methods and their respective interactions, but less pronounced

6.4. Variance Components, the Key to New Developments. Variance components add extra value to *random line* models. They are new in the context of school effectiveness research, so new in fact that, as far as we know, no theories exists that predict these parameters. So far they are considered mostly as a nuisance, and usually they are not reported. If they are reported they get the notion of residual variance, ready to be explained by second level variables. The same way as in linear regression, the sigma squared is considered a potential variance in the dependent variable that is there waiting to be explained. Only two studies are available that explore if variance components are estimated in an unbiased way, and if they differ from estimates based on weighted least squares (GLS), which are Busing [9] and Van der Leeden and Busing [36]. This last study shows that the variance components are estimated with increasing precision over iterations in ML3. The Mean Squared error (MSE) decreases with the number of iterations. MSE is defined as the average squared distance from the estimator to the true value of the parameter. The MSE becomes smaller after each iteration under all conditions and under both IGLS or RIGLS. The bias for the different parameters are also assessed. It shows that for GLS, IGLS, and RIGLS the variance components are underestimated or downward biased. Bias is eliminated in large datasets, with a large number of groups.

7. Conclusions and Discussions

Conclusions based on the simulation studies are regarding the fixed effects or gamma's, or the variance components. As far as the gamma's are concerned we conclude based on the available studies that OLS, GLS or EB/ML are equally good given that the standard errors are calculated in the correct way, using equation (5).

The variance components are negatively biased in both GLS, (restricted) RIGLS, and (unrestricted) IGLS.

Our conclusions are based on simulated data, with low (but realistic) intra class correlations of 0.10 and 0.20, we come to the following conclusions. The conclusions are, that the *random line* model is technically an improvement over the traditional multiple regression model, because it calculates the correct standard errors. If the researcher is only interested in gamma's and standard errors a traditional multiple regression (with a correct procedure for standard errors) is sufficient. In other words: the program could stop after the first iteration. If researchers have an interest in variance components, the iteration procedure improves the estimates, because the estimates are less biased after convergence is reached than they are after the first iteration (the GLS estimates).

Researchers interested in prediction for separate schools, the *random line* model is technically superior, because the empirical Bayes procedure improves the estimation of the parameters for separate schools (the posterior means) by borrowing of strength. Borrowing of strength is especially important in the presence of sparse data, or small groups. If groups are large (and no borrowing of strength is necessary we don't suggest to use the posterior means (see also Goldstein [17]), but advice to use separate models for separate schools. In general it seems to us a good suggestion to start with such separate models if the data allow to do so. Separate analysis per school allows the researcher to explore the data and the pattern in the data before fitting the *random line* model. This can be done by plotting the residuals against the second level variables, or plot the different values for the slopes against the second level variables. But in the presence of sparse data (small groups or not enough information within groups, see Rubin [32]), the *random line* model is a good solution for prediction per school.

Conceptually the *random line* model is an improvement because it makes us think in a different ways. We believe that it will enhance discussion regrading school effects. We doubt if the model will provide us with more advanced theoretical knowledge.

The limitation of our study is, that we can only make inferences to data that resemble the simulated data in the studies summarized. We don't know what will happen if the data are very or even extremely unbalanced. We also expect that higher intra correlation than are usually found in education (higher than 0.20) will show larger differences between estimation procedures. We think however that the studies reviewed have shown that *random line* models do serve a purpose, but will very often lead to the same conclusions as classical regression models.

REFERENCES

1. M. A. Aitkin and N. T. Longford, *Statistical modeling issues in school effectiveness studies*, Journal of the Royal Statistical Society A **149** (1986), 1–43.
2. M. A. Aitkin and R. Zuzovsky, *New paradigm for the analysis of hierarchically structured date in school effectiveness studies*, Paper presented at the Annual meeting of the AERA, San Francisco, April 20-24, 1992., 1992.
3. R. S. Barcikowski, *Satistical power with group mean as the unit of analysis*, Journal of Educational Statistics **6** (1981), 267–285.
4. B. Bassiri, *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model*, Paper presented at the annual meeting of the American Educational research Association. Chicago, University of Chicago, 1988.
5. R.D. Bock, *Multilevel analysis of educational data*, Academic Press, New York, NY, 1989.
6. A. S. Bryk and S. Raudenbush, *Hierachical linear models for social and behavioural research:applications and data analysis methods*, Sage Publications, Newbury Park, CA, 1991.
7. 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.
8. 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.
9. F. M. T. A. Busing, *Distribution characteristics of variance estimates in two-level models.*, Preprint PRM 93-04, Psychometrics and Research Methodology, Leiden, Netherlands, 1993.
10. L. J. Cronbach and R. E. Snow, *Aptitude and instructional methods*, Irvington, New York, NY, 1977.
11. L. J. Cronbach and N. Webb, *Between-class and within-class effects in a reported aptitude × treatment interaction: Reanalysis of a study by G.L. Anderson*, Journal of Educational Psychology **67** (1975), 717–724.
12. J. de Leeuw, *Data modeling and theory construction*, Operationalization and Research Strategy (Amsterdam, Netherlands) (J. Hox and L. de Jong-Gierveld, eds.), Swets and Zeitlinger, 1990.
13. J. de Leeuw and I.G.G. Kreft, *Random Coefficient Models for Multilevel Analysis*, Journal of Educational Statistics **11** (1986), 57–86.
14. D. Draper, *Inference and hierarchical modeling in the social sciences*, Paper presented at the workshop on Multilevel Analysis organized by RAND Corporation and UCLA Statistics, Santa Monica, October 3-4., 1993.
15. C. T. Fitz-Gibbon, *Multilevel Modelling in an Indicator System*, Schools, Classrooms and Pupils. International Studies of Schooling from a Multilevel Perspective (New York, NY) (S. Raudenbush and J. Willms, eds.), Academic Press, 1991.
16. H. Goldstein, *Multilevel models in educational and social research*, Griffin, London, GB, 1987.
17. ———. *Commentary: Better Ways to Compare Schools ?*, Journal of Educational Statistics **16** (1992), 89–92.
18. J. Gray, *Multilevel Models: Issues and Problems Emerging from their Recent Application in British Students of School Effectiveness*, Multilevel Analysis of Educational Data (New York, NY) (R. D. Bock, ed.), Academic Press, 1989.
19. K.-S. Kim, *Multilevel Data Analysis: a Comparison of Analytical Alternatives*, Ph.D. thesis, University of California, Los Angeles, 1990.
20. I. G. G. Kreft, *The analysis of small group data. a reanalysis of webb 1982 with a random coefficient model*, Paper presented at the Annual meeting of the AERA, San Francisco, April 20-24, 1992. Eric Document TMO 18787, 1992.
21. I. G. G. Kreft and J. de Leeuw, *Model-based Ranking of Schools*, International Journal of Edu-

cation **15** (1991), 45–59.
22. I. G. G. Kreft, J. de Leeuw, and R. van der Leeden, *Review of Five Multilevel Analysis Programs: BMDP-5V, GENMOD, HLM, ML3, VARCL*, American Statistician (in press).
23. 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.
24. N. T. Longford, *Variance Component Models and Observational Data*, Theory and Model in Multilevel Resarch: Convergence or Divergence? (Amsterdam, Netherlands) (P. van der Eeden, J. Hox, and J. Hauer, eds.), SISWO, 1990.
25. N.T. Longford, *VARCL. software for variance component analysis of data with nested random effects (maximum likelihood)*, Educationai Testing Service, Princeton, NJ, 1990.
26. W. M. Mason, G. Y. Wong, and B. Entwisle, *Contextual Analysis through the Multilevel Linear Model*, Sociological Methodology (1984), 72–103.
27. R. P. McDonald, *Some model for the bilevel bivariate relationship*, Paper presented at the workshop on Multilevel Analysis organized by RAND Corporation and UCLA Statistics, Santa Monica, October 3-4., 1993.
28. ———, *The Bilevel Reticular Action Model for Path Analysis With Latent Variables*, Sociological Methods and Research **22** (1994), 399–413.
29. B. O. Muthén, *Multilevel Covariance Structure Analysis*, Sociological Methods and Research **22** (1994), 376–399.
30. R. Prosser, J. Rabash, and H. Goldstein, *ML3, software for three-level analysis. Users guide for V2.*, Institute of Education, University of London, London, GB, 1990.
31. D. Rachman-Moore and R. G. Wolfe, *Robust Analysis of a Non-linear Model for Educational Survey Data*, Journal of Educational Statistics **9** (1984), 277–294.
32. D. B. Rubin, *Some applications of Multilevel Models of Educational Data*, Multilevel Analysis of Educational Data (New York, NY) (R. D. Bock, ed.), Academic Press, 1989.
33. T. A. Snijders and R. J. Bosker, *Standard Errors and Sampling Sizes for Two-level Research*, Journal of Educational Statistics **18** (1993), 237–261.
34. 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.
35. R. L. Tate and Y. Wongbundit, *Random versus Nonrandom Coefficient Models for Multilevel Analysis*, Journal of Educational Statistics **8** (1983), 103–120.
36. R. van der Leeden and F. M. T. A. Busing, *First iteration versus igls/rigls estimates in two-level models: a monte carlo study with ml3*, Preprint PRM 94-03, Psychometrics and Research Methodology, Leiden, Netherlands, 1994.
37. N. M. Webb, *Group Composition, Group Cooperation, and Achievement in Cooperative Small Groups*, Journal of Educational Psychology **74** (1982), 475–484.
38. G. Wunsch, *Causal theory and causal modeling*, Leuven University Press, Leuven, Belgium, 1988.

SCHOOL OF EDUCATION, DIVISION OF EDUCATIONAL FOUNDATIONS, CALIFORNIA STATE UNIVERSITY, 5151 STATE UNIVERSITY DRIVE, LOS ANGELES, CA 99032, TEL: 213 - 343 5116, FAX: 213 - 343 4318

E-mail address: kreft@laplace.stat.ucla.edu

RESEARCH DIVISION, CTB MCGRAW-HILL, 20 RYAN RANCH RD., MONTEREY, CA 93940

E-mail address: iao4byo@mvs.oac.ucla.edu