APPLIED MULTILEVEL ANALYSIS

J.J. Hox

TT-Publikaties, Amsterdam

1995

CIP-GEGEVENS KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Hox, J.J.

Applied multilevel analysis
/ J.J. Hox. - Amsterdam: TT-Publikaties.
- Ill., fig., tab. + diskette
- Index, lit.
ISBN 90-801073-2-8
NUGI 659
Trefw.: multilevel analyse

CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

... of making many books there is no end;
and much study is a weariness of the flesh.

Ecclesiastes 12:12

This book is meant as a basic and fairly nontechnical introduction to multilevel analysis, for applied researchers in the social sciences. The term 'multilevel' refers to a hierarchical or nested data structure, usually people within organizational groups, but the nesting may also consist of repeated measures within people, or respondents within clusters as in cluster sampling. The expression *Multilevel model* or *multilevel analysis* is used as a generic term for all models for nested data. This book presents two multilevel models: the multilevel regression model and a model for multilevel covariance structures.

I thank Rian van Blokland-Vogelesang, Pieter van den Eeden, Edith de Leeuw, Godfried van den Wittenboer, and Tom Snijders for their comments on earlier drafts. I also thank Jaap Dronkers, Edith de Leeuw, Arie van Peet, Bert Schijf and Kees van der Wolf for their permission to use their data for the examples.

I gratefully acknowledge the organizational support of my employing organization, the Faculty of Educational Sciences of the University of Amsterdam. Furthermore, I have enjoyed the opportunity to stay as a Fulbright scholar at the Department of Psychology and the Social Statistics Program of the University of California, Los Angeles. I thank both organizations for providing a stimulating research environment.

My research has benefited from the lively discussions in various research committees of which I am a member. I specifically want to mention the SOMO/NOSMO research committees on Multilevel Research (MULOG) and on Conceptualization and Research Design.

This second edition of Applied Multilevel Analysis follows the text of the first edition, with a few alterations. First, I have corrected a number of small errors in the text and the equations. Second, I have incorporated a number of improvements suggested in reviews by Ian Plewis and Rian van Blokland-Vogelesang. Third, I have rewritten part of the material on the programs HLM and MLn to reflect software upgrades. Still, the chapter on HLM, VARCL and MLn should not be read as a complete introduction to the software; the available program packages are much more powerful than I can show here.

Multilevel analysis is a complex and diverse field. The goal of this book is to provide an introduction to the basic approach and purpose. It can only give a first impression of the great diversity and depth of multilevel analysis. Several more advanced texts are available to researchers for further study. In addition, the Multilevel Models Project of The University of London publishes a Multilevel Modelling Newsletter (information available via the multilevel modeling website *http://www.ioe.ac.uk/multilevel/*>. There is also an e-mail distribution list for multilevel research (via mailbase@mailbase.ac.uk).

J.J. Hox

Amsterdam
September 1995

Notes on the electronic edition:

This is the complete text of the book 'Applied Multilevel Analysis.' The book is now out of print, and will not be reprinted because I feel it is becoming outdated. Since many people still consider it a very readable introduction to the basics of multilevel analysis, I have decided to make it available as an electronic web document. The usual copyright still applies; anyone may use this material for noncommercial purposes, provided the original source is referenced and fully cited.

To turn the original file into a PDF file, I have converted to a Windows program. As a result, the page format has changed slightly, and the page numbers in the index may not be completely accurate. If readers send me corrections, I will include these in later electronic editions.

Amsterdam
January 1999

1. Introduction

Social research often concerns problems that investigate the relationship between individual and society. The general concept is that individuals interact with the social contexts to which they belong, meaning that individual persons are influenced by the social groups to which they belong, and that the properties of those groups are in turn influenced by the individuals who make up that group. Generally the individuals and the social groups are conceptualized as a hierarchical system of individuals and groups, with individuals and groups defined at separate levels of this hierarchical system. Naturally, such systems can be observed at different hierarchical levels, and as a result may have variables defined at each level. This leads to research into the interaction between variables that describe the individuals and variables that describe the social groups, a kind of research that is now often referred to as `*multilevel research*'.

In multilevel research, the data structure in the population is hierarchical, and the sample data are viewed as a multistage sample from this hierarchical population. Thus, in educational research, the population consists of schools and pupils within these schools, and the sampling procedure proceeds in two stages: first we take a sample of schools, and next we take a sample of pupils within each school.[1] In this example, pupils are said to be nested within schools. Other examples are cross-national studies where the individuals are nested within their national units, organizational research with individuals nested within organizations, family research with family members within families, and methodological research into interviewer effects with respondents nested within interviewers. Less obvious applications of multilevel models are longitudinal research and growth research where several distinct observations are nested within individuals, and meta-analysis where the subjects are nested within different studies. For simplicity I will mostly describe the multilevel models in this

[1]Of course in real research one may have a convenience sample at either level, or one may decide not to sample pupils but to study all pupils in the sample of schools. Nevertheless, one should keep firmly in mind that the central notion is one of successive sampling from each level of a hierarchical population.

book in terms of individuals nested within groups, and use examples about pupils nested within schools, but note that the models apply to a much larger class of analysis problems.

In multilevel research, variables can be defined at any level of the hierarchy. Some of these variables may be measured directly at their natural level; for example, at the school level we may measure school size and denomination, and at the pupil level intelligence and school success. In addition, we may move variables from one level to another by aggregation or disaggregation. Aggregation means that the variables at a lower level are moved to a higher level, for instance by computing the school mean of the pupils' intelligence scores. Disaggregation means moving variables to a lower level, for instance by assigning to all pupils a variable that reflects the denomination of the school they belong to. Lazarsfeld and Menzel (1961) give a typology to describe the relations between different types of variables, defined at different levels. I present them in the following scheme, adapted from Swanborn (1981):

Level:	1		2		3	et cetera
Variable	absolute	⇒	analytical			
type:	relational	⇒	structural			
	contextual	⇐	global	⇒	analytical	
			relational	⇒	structural	
			contextual	⇐	global	⇒
					relational	⇒
					contextual	⇐

In this scheme, the lowest level (level 1) is usually formed by the individuals. However, this is not always the case. Galtung (1969), for instance, defines roles within individuals as the lowest level, and in longitudinal designs one can define repeated measures within individuals as the lowest level (Goldstein, 1986, 1989.) At each level, we have several types of variables. *Global* and *absolute* variables

refer only to the level at which they are defined, without reference to any other units or levels ('absolute variables' is simply the term used for global variables defined at the lowest level). A pupil's intelligence would be a global or absolute variable. *Relational* variables also refer to one single level, they describe the relationships of a unit to the other units at the same level. Many sociometric indices (such as indices of popularity of indices of the reciprocality of relationships) are relational variables. *Analytical* and *structural* variables are measured by referring to the subunits at a lower level. Analytical variables refer to the distribution of an absolute or a global variable at a lower level, for instance to the mean of a global variable from a lower level. Structural variables refer to the distribution of relational variables at the lower level; many social network indices are of this type. Constructing an analytical or relational variable from the lower level data involves *aggregation* (indicated by $\Rightarrow$): data on lower level units are aggregated into data on a smaller number of higher level units. *Contextual* variables, on the other hand, refer to the superunits; all units at the lower level receive the value of a variable for the superunit to which they belong at the higher level. This is called *disaggregation* (indicated by $\Leftarrow$): data on higher level units are disaggregated into data on a larger number of lower level units. The resulting variable is called a *contextual* variable, because it refers to the higher level context of the units we are investigating.

For the purpose of analyzing multilevel models, it is usually not important to assign each variable its proper place in the scheme given above. The advantage is conceptual; the scheme makes clear to which level the measurements properly belong. Historically, multilevel problems have led to analysis approaches that move all variables by aggregation or disaggregation to one single level of interest, followed by an ordinary multiple regression, analysis of variance, or some other `standard' analysis method. For example, an explicitly multilevel or contextual theory in education is the so-called `frog pond' theory, which refers to the idea that a specific individual frog may either be a small frog in a large pond or a large frog in a small pond. Applied to education, this metaphor points out that the effect of an explanatory variable such as `intelligence' on school career may depend a lot on the average intelligence in the school. A moderately intelligent pupil in a highly intelligent context may become demotivated and thus become an underachiever,

while the same pupil in a considerably less intelligent context may gain confidence and become an overachiever. Thus, the effect of an individual pupil's intelligence depends on the average intelligence of the other pupils. A popular approach in educational research to investigate `frog pond' effects has been to aggregate variables into group means, and then to disaggregate these group means again to the individual level. As a result, the data file contains both individual level (absolute or global) variables and higher level (contextual) variables in the form of the disaggregated group means. Cronbach (1976; cf. Cronbach & Webb, 1979) has suggested to express the individual scores as deviations from their respective group means, a procedure that has become known as *centering around the group mean*, or *group centering*. Centering around the group means makes very explicit that the individual scores should be interpreted relative to their group's mean. Another advantage of centering around the group means is that the group-centered individual deviation scores have a zero correlation with the disaggregated group means, which has statistical advantages. However, a definite disadvantage of centering around the group means is that the group-centered variables have no longer a simple interpretation. I will not go into the problem of centering here; for a thorough discussion of the conceptual and analytical implications of various centering schemes I refer to Boyd and Iversen (1979), Iversen (1991), and the discussion by Raudenbush (1989a, 1989b), Longford (1989b) and Plewis (1989). However, the discussion of the `centering' issue makes clear that combining and analyzing information from different levels within one statistical model is central to multilevel modeling.

Analyzing variables from different levels at one single common level creates two different sets of problems. One set of problems is statistical. If data are aggregated, the result is that different data values from many subunits are combined into fewer values for fewer higher level units. Information is lost, and the statistical analysis looses power. On the other hand, if data are disaggregated, the result is that a few data values from a small number of superunits are `blown up' into values for a much larger number of subunits. Ordinary statistical tests treat all these disaggregated data values as independent information from this much larger sample. The proper sample size for these variables is of course the number of higher level units. Using the higher number of disaggregated cases for

the sample size leads to significance tests that reject the null-hypothesis far more often than the nominal alpha level suggests. In other words: investigators come up with a lot of spurious significances.

The other set of problems encountered is conceptual. If the analyst is not very careful in the interpretation of the results, s/he may commit the fallacy of the wrong level, which consists of analyzing the data at one level, and drawing conclusions at another level. Probably the best known fallacy is the *ecological fallacy*, which is interpreting aggregated data at the individual level. It is also know as the `Robinson effect' after Robinson (1950). Robinson presents aggregated data describing the relationship between the percentage of blacks and the illiteracy level in nine geographic regions in 1930. The *ecological correlation* (correlation between the aggregated variables) at the region level is 0.95, but the individual correlation between the absolute variables at the individual level is 0.20. Robinson concludes that in practice an ecological correlation is almost certainly not equal to its corresponding individual correlation. This has consequences the other way as well; drawing inferences at a higher level from analyses performed at a lower level is just as misleading; this error is known as the *atomistic fallacy*. An extensive typology of such fallacies is given by Alker (1969). A different but related fallacy is known as `Simpson's Paradox' (see Lindley & Novick, 1981). Simpson's paradox refers to the problem that completely erroneous conclusions may be drawn if grouped data, drawn from heterogeneous populations, are collapsed and analyzed as if they came from a single homogeneous population.

A more general way to look at multilevel data is to investigate cross level hypotheses, or *multilevel* problems. A multilevel problem is a problem that concerns the relationships between variables that are measured at a number of different hierarchical levels. For example, a common question is how a number of individual and group variables influence one single individual outcome variable. Typically, some of the higher level explanatory variables may be the aggregated group means of lower level individual variables. The goal of the analysis is to determine the direct effect of individual and group level explanatory variables, and to determine if the explanatory variables at the group level serve as moderators of individual-level relationships. If group level variables moderate

lower level relationships, this shows up as a statistical interaction between explanatory variables from different levels. In the past, such data were usually analyzed using conventional multiple regression analysis with one dependent variable at the lowest (individual) level and a collection of explanatory variables from all available levels (Boyd & Iversen, 1979; Roberts & Burstein, 1980; Van den Eeden and Hüttner, 1982). Since this approach analyzes all available data at one single level, it suffers from the conceptual and statistical problems mentioned above. Much research has been directed at developing more appropriate analysis methods for this hierarchical regression model, and at clarifying the associated conceptual and statistical issues.

1.1. Why Do We Need Special Multilevel Analysis Techniques?

A multilevel problem concerns a population with a hierarchical structure. A sample from such a population can be described as a multistage sample: first we take a sample of units from the higher level (e.g., schools), and next we sample the subunits from the available units (e.g., we sample pupils from the schools). In such samples, the individual observations are generally not completely independent. For instance, pupils in the same school tend to be similar to each other, because of selection processes (e.g., some schools may attract primarily higher SES pupils, while others attract more lower SES pupils) and because of the common history they share by going to the same school. As a result, the average correlation (expressed in the so-called *intra class correlation*) between variables measured on pupils from the same school will be higher than the average correlation between variables measured on pupils from different schools. Standard statistical tests lean heavily on the assumption of independence of the observations. If this assumption is violated (and in multilevel data this is usually the case) the estimates of the standard errors of conventional statistical tests are much too small, and this results in many spuriously `significant' results.

The problem of dependencies between individual observations also occurs in survey research, when the sample is not taken at random but cluster sampling

from geographical areas is used instead. For similar reasons as in the school example given above, respondents from the same geographical area will be more similar to each other than respondents from different geographical areas. The result is again estimates for standard errors that are too small, and spurious `significant' results. In survey research this is called a `design effect', and the usual correction procedure is to compute the standard errors by ordinary analysis methods, estimate the intra class correlation between respondents within clusters, and to employ a correction formula to the standard errors (cf. Kish, 1987). Some of these correction procedures are quite powerful (cf. Skinner, Holt & Smith, 1989). As a matter of fact, these correction procedures could also be applied in multilevel analysis. However, in most multilevel problems we have not only clustering of individuals within groups, but we also have variables measured at all available levels. Combining variables from different levels in one statistical model is a different problem than estimating and correcting for design effects. Multilevel models are designed to analyze variables from different levels simultaneously, using a statistical model that includes the various dependencies.

1.2. Multilevel Theories

Multilevel problems must be explained by multilevel theories, an area that seems underdeveloped compared to the advances made in the recently developed modeling and computing machineries (cf. Van den Eeden, 1993). If there are effects of the social context on individuals, these effects must be mediated by intervening processes that depend on characteristics of the social context. Multilevel models so far require that the grouping criterion is clear, and that variables can be assigned unequivocally to their appropriate level. In reality, group boundaries are sometimes fuzzy and somewhat arbitrary, and the assignment of variables is not always obvious and simple. In multilevel problems, decisions about group membership and operationalizations involve a wide range of theoretical assumptions, and an equally wide range of specification problems for the auxiliary theory (cf. Blalock, 1990). When the numbers of variables at different levels are large, there is an enormous number of possible cross-level interactions.

Ideally, a multilevel theory should specify which variables belong to which level, and which direct effects and cross-level interaction effects can be expected. Cross-level interaction effects between the individual and the context level require the specification of some process within individuals that causes those individuals to be differentially influenced by certain aspects of the context. Attempts to identify such processes have been made by, among others, Stinchcombe (1967), Hummel (1972), and Erbring and Young (1979). The common denominator in these theories is that they all postulate one or more psychological processes that mediate between individual variables and group variables. Since a global explanation by `group telepathy' is generally not acceptable, communication processes and the internal structure of groups become important. This refers to the `structural variables' mentioned earlier. In spite of their theoretical relevance, structural variables are infrequently used in multilevel research (the program developed by Erbring and Young to include sociometric structures in multilevel analysis is also seldom used, possibly because it is not readily accessible to applied researchers.) Another theoretical area that has been largely neglected by multilevel researchers is the influence of individuals on the group. This is already visible in Durkheim's concept of sociology as a science that focuses primarily on the constraints that a society can put on its members, and disregards the influence of individuals on their society.

1.3. Models Described in This Book

This book treats two classes of multilevel models: multilevel regression models, and multilevel models for covariance structures.

Multilevel regression models are essentially a multilevel version of the familiar multiple regression model. As Cohen and Cohen (1983) and others have shown, the multiple regression model is very versatile. Using dummy coding for categorical variables, it can be used to analyze analysis of variance (ANOVA)-type of models as well as the more usual multiple regression models. As a result the multilevel regression model can be used in a wide variety of research problems. It has been used extensively in educational research (cf. the special issues of the International Journal of Educational Research, 1990, the Dutch Tijdschrift voor

Onderwijsresearch also in 1990, and Journal of Educational and Behavioral Statistics in 1995). Other applications have been in the analysis of longitudinal and growth data (cf. Bryk & Raudenbush, 1987; Goldstein, 1989; DiPrete & Grusky, 1990, Goldstein, Healy & Rasbash, 1994), the analysis of interview survey data (cf. Hox, de Leeuw & Kreft, 1991; Hox, 1994), data from surveys with complex sampling schemes with respondents nested within sampling units (Goldstein & Silver, 1989), and data from factorial surveys and facet designs (Hox, Kreft & Hermkens, 1991; Hox & Lagerweij, 1993). Raudenbush and Bryk have introduced multilevel regression models in meta-analysis (cf. Raudenbush & Bryk, 1985, 1987; Hox & de Leeuw, 1994). Multilevel regression models for binary and other non-normal data have been described by Wong and Mason (1985), Longford (1988), Mislevy and Bock (1989) and Goldstein (1990). This book describes the multilevel version of the usual multiple regression model at length in chapter 2, and an extended example of analyses with the programs HLM, MLn and VARCL in chapter 3. Chapter 4 gives examples of some special applications, such as analysis of proportions and meta-analysis.

Multilevel covariance structure analysis (CSA) would constitute a very powerful tool for the analysis of multilevel data. Quite a lot of fundamental work has been done on multilevel factor and path analysis (cf. Goldstein and McDonald, 1988; Muthén, 1989, 1990; McDonald & Goldstein, 1989). There are also some applications, for instance Härnqvist, Gustaffson, Muthén, and Nelson (1992), Hox (1993). However, as yet there is almost no specialized software to analyze multilevel covariance structures; applications currently require using experimental programs (such as BIRAM, McDonald, 1994) or running conventional software for covariance structure analysis (e.g., Lisrel, EQS, Liscomp) with unusual setups. The general statistical model for multilevel covariance structure analysis is quite complicated. Chapter 5 in this book describes a simplified statistical model proposed by Muthén (1990, 1994), and explains how multilevel confirmatory factor and path models can be estimated with conventional CSA software such as Lisrel.

2. Multilevel Regression Models

2.1 The Basic Two-Level Regression Model

The multilevel regression model has become known in the research literature under a variety of names, such as `random coefficient model' (de Leeuw & Kreft, 1986; Longford, 1993), `variance component model' (Longford, 1986), hierarchical linear model' (Raudenbush & Bryk, 1986, 1992). The models described in these publications are not *exactly* the same (especially when computational details are considered,) but they are highly similar, and I will refer to them collectively as `multilevel regression models'.

The full multilevel regression model assumes that there is a hierarchical data set, with one single dependent variable that is measured at the lowest level and explanatory variables at all existing levels. Conceptually the model can be viewed as a hierarchical system of regression equations. For example: assume that we have collected data in *J* schools, with data from a different number of pupils N_j in each school. On the pupil level we have the dependent variable `school career outcome' (Y) and the explanatory variable `SES' (X), and on the school level we have the explanatory variable `school size' (Z). Accordingly, we can set up a separate regression equation in each separate school to predict the dependent variable Y by the explanatory variable X as follows:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij}. \qquad (2.1)$$

In this regression equation β_{0j} is the usual intercept, β_{1j} is the usual regression coefficient (regression slope), and e_{ij} is the usual residual error term. The subscript *j* is for the schools (j=1..J) and the subscript *i* is for individual pupils (i=1..N_j). The difference with the usual regression model is that we assume that each school is characterized by a different intercept coefficient β_{0j} and also a different slope coefficient β_{1j}. Just as in the ordinary multiple regression model, the random errors e_{ij} in each school are assumed to have a mean of zero and a variance which is specified as σ_j^2; most multilevel models simply assume that the random error

variance is the same in all schools and specify this common error variance as σ^2.

In other words: the intercept and slope coefficients are assumed to vary across the schools; for that reason they are often referred to as *random* coefficients.[1] In our example each school is characterized by its own specific value for the intercept and the slope coefficient for the pupil variable `SES'. For pupils with the same score on the explanatory variable SES, a school with a high value of the intercept is predicted to lead to a higher school career outcome than a school with a low value for the intercept. Similarly, the differences in the values for the slope coefficient for SES can be interpreted to mean that the relationship between the social background of the pupils and their predicted career is not the same in all schools. Some schools have a high value for the slope coefficient of SES; in these schools SES has a large effect on the school career and we might describe these schools as `selective'. Other schools have a low value for the slope coefficient of SES; in these schools SES has a small effect on the school career and we could describe these schools as `egalitarian'.

Across all schools, the regression coefficients β_j have a distribution with some mean and variance. The next step in the hierarchical regression model is to predict the variation of the regression coefficients β_j by introducing explanatory variables at the school level, as follows:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \qquad (2.2)$$

and

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \qquad (2.3)$$

[1]Of course they are not assumed to be *completely* random. We hope to be able to explain at least some of this variation by introducing higher level variables. However, in most cases we will not be able to explain all this variation, and as a result after introducing the higher level variables there will be some random variation left unexplained. Hence the name 'random coefficient model' for this type of model: the regression coefficients (intercept and slopes) are assumed to have some amount of random variation between schools. The name 'random component model' refers to the statistical problem of estimating the amount of this random variation.

Equation (2.2) states that the general career level of a school (the intercept β_{0j}) can be predicted by the school size (Z). Thus, if $ß_{01}$ is positive, we state that the school career outcome in large schools is higher than in small schools. Conversely, if $ß_{01}$ is negative, we state that the school career outcome in large schools is lower than in small schools. The interpretation of equation (2.3) is more complicated. Equation (2.3) states that the *relationship* (as expressed by the slope coefficient $ß_{1j}$) between the school career (Y) and the SES (X) of the pupil depends upon the school size (Z). Whether a school is `selective' (high value for $\beta_{1j}$) or `egalitarian' (low value for $ß_{1j}$), depends (at least partly) upon the school's size. If γ_{11} is positive, large schools tend to be more selective than small schools, and if γ_{11} is negative, large schools are more egalitarian than small schools. Thus, the school size acts as a *moderator variable* for the relationship between school career and SES; this relationship varies according to the value of the moderator variable. For a statistical discussion on how to interpret differences between regression equations for different schools see Aitkin and Longford (1986); more application oriented discussions are offered in Kreft and de Leeuw (1991, 1993).

The u-terms u_{0j} and u_{1j} in equations (2.2) and (2.3) are (random) residual error terms at the school level. The residual errors $u_{.j}$ are assumed to have a mean of zero, and to be independent from the residual errors e_{ij} at the individual (pupil) level. The variance of the residual errors u_{0j} is specified as σ_{00}, and the variance of the residual errors u_{1j} is specified as σ_{11}. The *covariance* σ_{12} between the residual error terms u_{0j} and u_{1j} is generally not assumed to be zero.

Note that in equations (2.2) and (2.3) the regression coefficients γ are not assumed to vary across schools (consequently they have no subscript *j* to indicate to which school they belong: they apply to *all* schools). Therefore they are referred to as *fixed* coefficients, all between school variation left in the β coefficients after predicting these with the school variable Z_j is assumed to be residual error variation, which is captured by the residual error terms u_j (which therefore do have subscripts *j* to indicate to which school they belong).

Our model with one pupil level and one school level explanatory variable can be written as one single complex regression equation by substituting equations (2.2) and (2.3) into equation (2.1). Rearranging terms gives:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij} \qquad (2.4)$$

The segment $\gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij}$ in equation (2.4) contains all the fixed coefficients; for this reason this is often called the fixed (or deterministic) part of the model. The segment $u_{0j} + u_{1j} X_{ij} + e_{ij}$ in equation (2.4) contains all the random error terms; for this reason this is often called the random (or stochastic) part of the model. The term $Z_j X_{ij}$ is an interaction term that appears in the model as a consequence of modeling the varying regression slope β_{1j} of pupil level variable X_{ij} with the school level variable Z_j. Thus, the moderator effect of Z on the relationship between the dependent variable Y and X is expressed as a *cross-level interaction*. The interpretation of interaction terms in multiple regression analysis can be complex. In general, the substantive interpretation of the coefficients in models with interactions is much simpler if the variables that make up the interaction are expressed as deviations from their respective means. (Both the overall mean as the group means will do. Since centering around the group means introduces its own set of problems, I generally prefer to center around the overall mean. For a discussion see Raudenbush 1989a, 1989b; Longford, 1989b, Plewis, 1989.) I present an example of a cross-level interaction in section 2.5; for a thorough discussion of interactions in multiple regression models see Jaccard, Turrisi and Wan (1990) and Aiken and West (1991). Note that the random error term u_{1j} is connected to X_{ij}. Since the error term u_{1j} is multiplied by the explanatory variable X_{ij}, the resulting total error will be different for different values of X_{ij}, a situation which in ordinary multiple regression is called `heteroscedasticity'.[1]

As I explained in the introduction in chapter 1, multilevel models are needed because with grouped data the observations in the same group are generally more similar than the observations from different groups, which violates the assumption of independence of all observations. This lack of independence can be expressed as

[1]Put the other way around: the usual multiple regression model assumes 'homoscedasticity', meaning that all the errors are independent of all explanatory variables. If this assumption is not true, ordinary multiple regression does not work very well, which is one reason why analyzing multilevel models with ordinary multiple regression programs does not work very well either.

a correlation coefficient: the intra class correlation. The methodological literature contains a number of different formula's to estimate the intra class correlation r. For example, if we use oneway analysis of variance to test if there is a significant group effect, the intra class correlation is estimated by r = (MS(A)-MS(error))/(MS(A)-(k-1)xMS(error)). Shrout and Fleiss (1979) give an overview of other anova based formula's for different designs. The multilevel regression model can also be used to estimate the intra class correlation. The model used for this purpose is a model that contains no explanatory variables at all, the so-called *intercept-only* model. This can be derived from equations (2.1) and (2.2) as follows. If there are *no* explanatory variables X at the lowest level, equation (2.1) reduces to:

$$Y_{ij} = \beta_{0j} + e_{ij}. \qquad (2.5)$$

Likewise, if there are no explanatory variables Z at the highest level, equation (2.2) reduces to:

$$\beta_{0j} = \gamma_{00} + u_{0j}. \qquad (2.6)$$

We find the single equation model by substituting (2.6) into (2.5):

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \qquad (2.7)$$

We could also have found equation (2.7) by simplifying equation (2.4), removing all terms that contain an X or Z variable. The model of equation (2.7) does not explain any variance, it only decomposes the variance into two independent components: σ^2, which is the variance of the lowest level errors e_{ij}, and σ_{00}, which is the variance of the highest level errors u_{0j}. Using this model we can estimate the intra class correlation r by the equation:

$$r = \sigma_{00} / (\sigma_{00} + \sigma^2). \qquad (2.8)$$

The intra class correlation r is a population estimate of the variance explained by

the grouping structure. Equation (2.8) simply states that the intra class correlation is equal to the estimated proportion of group level variance compared to the estimated total variance. (Note that the intra class correlation is an estimate of the proportion of explained variance in the population. The amount of explained variance in the sample is the correlation ratio η^2 (eta-squared), cf. Hays, 1973).

2.2 Computing Parameter Estimates and Analysis Strategy

In general there will be more than one explanatory variable at the lowest level and also more than one explanatory variable at the highest level. Assume that we have P explanatory variables X at the lowest level, indicated by the subscript p (p=1..P). Likewise, we have Q explanatory variables Z at the highest level, indicated by the subscript q (q=1..Q). Then, equation (2.4) becomes the more general equation:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} Z_{qj} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} \qquad (2.9)$$

The errors at the lowest level e_{ij} are assumed to have a normal distribution with a mean of zero and a common variance σ^2 in all groups. The u-terms u_{0j} and u_{pj} in equation (2.9) are the error terms at the highest level. They are assumed to be independent from the errors e_{ij} at the individual level, and to have a multivariate normal distribution with means of zero. The variance of the residual error u_{0j} is the variance of the intercepts between the groups; it is specified as σ_{00}. The variances of the residual errors u_{pj} are the variances of the slopes between the groups; they are specified as σ_{pp}. The *covariances* between the residual error terms $\sigma_{p'p''}$ are generally not assumed to be zero; they are collected in the higher level variance/covariance matrix Σ.[1]

The statistical theory behind the multilevel regression model is complex. On the basis of the observed data, we want to estimate the parameters of the multilevel regression model: the regression coefficients and the variance components. The

[1]We may attach a subscript to Σ to indicate to which level it belongs. As long as there is no risk of confusion, I will use the simpler notation without the subscript.

estimators currently used in multilevel regression analysis are Maximum Likelihood (ML) estimators. Maximum Likelihood estimators estimate the parameters of a model by providing estimates for the population values that maximize the so-called Likelihood Function: the function that gives the probability of observing the sample data, given the current parameter estimates. To put it simply, ML estimates are those parameter estimates that maximize the probability of finding the sample data that we have actually found.

Maximum Likelihood procedures produce standard errors for most of the estimates. The standard errors can be used for significance testing; the test statistic *Z = parameter / (st.error param.)* can be referred to the standard normal distribution to establish a p-value for the null-hypothesis that in the population that specific parameter is zero. This test is known as the Wald test (Wald, 1943). The standard errors are asymptotic, i.e. they are valid for large samples. As usual, it is not precisely known when a sample is large enough to be confident about the precision of the estimates.[1] In ordinary regression analysis, a common rule of thumb is to require ten observations for each regression coefficient that is estimated. In multilevel regression, we should remember that higher level coefficients and variance components are estimated on the sample of groups, which is often not very large. Procedures for power analysis and some suggestions for decisions concerning sample sizes are given by Snijders and Bosker (1993).

It should be noted that the *p*-values produced by HLM may differ from those obtained from other programs. Most multilevel analysis programs produce as part of their output parameter estimates and asymptotic standard errors for these estimates, all obtained from the maximum likelihood estimation procedure. The usual significance test in a maximum likelihood is the Wald test: a *Z*-test of the form *Z=(estimate)/(standard error of estimate)* where *Z* is refered to the standard normal distribution. Bryk and Raudenbush (1992, p. 50), refering to a simulation study by Fotiu (1989), argue that for the fixed effects it is better to refer this ratio

[1]Since the standard errors are asymptotic, the p-values are in practical situations always an approximation. Given this fact, I prefer presenting standard errors rather than 'exact' p-values, and interpreting as statistically 'significant' those estimates that exceed two times their standard error. If finer distinctions are desired, estimates that exceed three times their standard error can be labeled as 'highly significant.'

to a *t*-distribution on *J-q-1* degrees of freedom. Likewise, they argue that the *Z*-test is not appropriate for the variances, because the sampling distribution of variances is skewed (Bryk and Raudenbush, 1992, p. 47, 55). Instead, they propose to use a chi-square test of the residuals. The *p*-values produced by HLM are based on these tests rather than the more usual Wald tests. When the number of groups *J* is small, the difference with the usual procedure may be important.

The Maximum Likelihood procedure also produces a statistic called the *deviance*, which indicates how well the model fits the data. In general, models with a lower deviance fit better than models with a higher deviance. If two models are *nested* (which means that a specific model can be derived from a more general model by removing parameters from the general model) the difference of the deviances for the two models has a chi-square distribution with degrees of freedom equal to the difference in the number of parameters estimated in the two models. This can be used to perform a formal chi-square test to test whether the more general model fits significantly better than the simpler model. The chi-square test of the deviances can also be used to good effect to explore the importance of random effects, by comparing a model that contains these effects with a model that excludes them. If the models to be compared are not nested models, the principle that models should be as simple as possible (theories and models should be parsimonious) indicates that we should generally stick with the simpler model.

Two different varieties of Maximum Likelihood estimation are currently used in the available software for multilevel regression analysis. One is called Full Maximum Likelihood (FML); in this method both the regression coefficients and the variance components are included in the likelihood function. The other method is called Restricted Maximum Likelihood (RML), here only the variance components are included in the likelihood function. The difference is that FML treats the estimates for the regression coefficients as known quantities when the variance components are estimated, while RML treats them as estimates that carry some amount of uncertainty (Bryk and raudenbush, 1992; Goldstein, 1995). Since RML is more realistic, it should, in theory, lead to better estimates, especially when the number of groups is small (Bryk & Raudenbush, 1992). However, in practice, the differences between the two methods are not very large (cf. Kreft, De Leeuw & Kim, 1989). FML has two advantages over RML: the

computations are generally easier, and since the regression coefficients are included in the likelihood function, the overall chi-square test can be used to test for differences between two models that differ only in the fixed part (the regression coefficients). With RML only differences in the random part (the variance components) can be tested with the overall chi-square test.

Computing the Maximum Likelihood estimates requires an *iterative* procedure. At the beginning the computer program generates reasonable starting values for the various parameters (in multilevel regression analysis these are usually based on the simple single-level regressions). In the next step, an ingenious computation procedure tries to improve upon the starting values, and produces better estimates. This second step is repeated (iterated) many times. After each iteration, the program inspects how much the estimates actually changed compared to the previous step. If the changes are very small, the program concludes that the estimation procedure has *converged* and that it is finished. Using these programs, we generally take the computational details for granted. However, sometimes computational problems do occur. A problem common to programs using an iterative Maximum Likelihood procedure is that the iterative process is not *guaranteed* to stop. There are models and data sets for which the program goes through an endless sequence of iterations, which can only be stopped by reaching for the <reset> switch on the computer. Because of this, most programs set a built-in limit to the maximum number of iterations. If convergence is not reached within this limit, the computations can be repeated with a higher limit. If the computations do not converge after a large number of iterations, we suspect that they may never converge.[1] The problem is how one should interpret a model that does not converge. The usual interpretation is that a model for which convergence cannot be reached is a bad model, using the simple argument that if estimates cannot be found this disqualifies the model. But the problem may also lie with the data. Especially with small samples the estimation procedure may fail even if the model is valid. Also, it is possible that, if only we had a better computation procedure, we could find acceptable estimates. Still, experience shows

[1]Some programs allow the analyst to monitor what happens during iterations, so one can observe whether the computations seem to be going somewhere, or are just moving back and forth without improving the likelihood function.

that if a program does not converge with a data set of reasonable size, the problem often is a badly misspecified model. In multilevel regression, nonconvergence often occurs when we try to estimate too many random (variance) components that are actually close or equal to zero. The solution is to simplify the model by leaving out some random components (generally the results from the non-converged solution provide an indication which random components can be omitted.)

The available multilevel regression programs all produce estimates for the fixed coefficients gamma, their standard errors, estimates for the variance components σ^2 and $\sigma_{p'p''}$, their standard errors, and the deviance. In addition, the various programs offer more analysis options. Some of the options offered by the programs HLM, VARCL and MLn will be compared in the next chapter.

Note that the number of parameters in a multilevel model is rather large. If there are *P* explanatory variables at the lowest level and *Q* explanatory variables at the highest level, the number of estimated parameters in the full model implied by equation (2.9) is given by the following list:

parameters:	number:
intercept	1
lowest level error variance	1
slopes for the lowest level predictors	P
highest level error variances for these slopes	P
highest level covariances of the intercept with all slopes	P
highest level covariances between all slopes	P(P-1)/2
slopes for the highest level predictors	Q
slopes for cross level interactions	P×Q

The ordinary single level regression model would estimate only the intercept, one error variance, and P+Q regression slopes. Clearly, even with a modest number of explanatory variables at both levels, equation (2.9) implies a very complicated model. Usually, we do not want to estimate the complete model, because this is likely to get us into computational problems, and also because it is very difficult to interpret such a complex model. Fortunately, we do not have to estimate the complete model. All programs allow us to specify which regression coefficients are

assumed to vary and which not, and to include only a few selected cross level interactions. So, generally we will limit ourselves to parameters that have proven their worth in previous research, or are interesting in view of our theoretical problem.[1]

If we have no strong theories, we can use an exploratory procedure to select a model. An attractive procedure is to start with the simplest possible model, the intercept-only model, and to include the various types of parameters step by step. At each step, we inspect the results to see which parameters are significant, and how much residual error is left at the two distinct levels. The different steps of such a selection procedure are given below.

Step 1:
Analyze a model with no explanatory variables. This model, the *intercept-only model*, is given by the model of equation (2.7), which is repeated here:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \; . \qquad (2.7)$$

The intercept-only model is useful because it gives us an estimate of the intra-class correlation by applying equation (2.8), which is repeated here:

$$r = \sigma_{00} / (\sigma_{00} + \sigma^2) \; . \qquad (2.8)$$

The intercept-only model also gives us the value of the deviance, which is a measure of the degree of mis-fit of the model (cf. McCullagh & Nelder, 1989).

Step 2:
Analyze a model with all lower level explanatory variables fixed. This means that the corresponding variance components of the slopes are fixed at zero. This model is written as:

---

[1]VARCL provides the option to restrict all covariances between random slopes to zero. (In HLM and MLn this can be accomplished by restricting them all one by one). This simplifies the model, and speeds up computations. In general, it is advisable to test the assumption that these covariances are zero by comparing the deviances of the models with and without covariances, which can be tested formally using a chi-square test with P(P-1)/2 degrees of freedom.

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij} . \qquad (2.10)$$

In this step we assess the contribution of each individual explanatory variable. If we use the FML estimation method, we can test the improvement of the final model chosen in this step by computing the difference of the deviance of this model and the previous model (the intercept-only model). This difference approximates a chi-square variate with as degrees of freedom the difference in the number of parameters of both models. In this case the degrees of freedom simply equal the number of explanatory variables added in step 2.

Step 3:
Assess whether any of the slope of any of the explanatory variables has a significant variance component between the groups. The model to consider is:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} . \qquad (2.11)$$

Testing random slope variation is best done on a one-by-one basis. Variables that were omitted at the previous step may be analyzed again at this step: it is quite possible for an explanatory variable to have no significant mean regression slope (as tested in step 2) but to have a significant variance component for this slope. After deciding which slopes have a significant variance between groups, we can add all the variance components in a final model and use the chi-square test based on the deviances to test whether the final model of step 3 fits better than the final model of step 2. (Since we are now introducing changes in the random part of the model, the chi-square test can also be used with RML estimation. When counting the number of parameters added, remember that step 3 also includes the covariances between the slopes!)

Step 4:
Add the higher level explanatory variables, as in equation (2.12):

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{pj} X_{pij} + u_{0j} + e_{ij} . \qquad (2.12)$$

This allows us to examine whether these variables explain between group variation in the dependent variable. Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit.

Step 5:
Add cross-level interactions between explanatory group level variables and those individual level explanatory variables that had significant slope variation in step 3. This leads to the full model already formulated in equation (2.9):

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} Z_{qj} X_{pij} + u_{pj} X_{pij} + u_{0j} + e_{ij} \qquad (2.9)$$

Again, if we use FML estimation, we can use the global chi-square test to formally test the improvement of fit.

In each step, we decide which regression coefficients or (co)variances to keep on the basis of the significance tests, the change in the deviance, and changes in the variance components. Specifically, if we introduce explanatory variables in step 2, we expect that the lowest level variance σ^2 goes down. If the composition of the groups with respect to the explanatory variables is not exactly identical for all groups, we expect that the higher level variance σ_{00} also goes down. Thus, the individual level explanatory variables explain part of the individual and part of the group variance. The higher level explanatory variables added in step 4 can explain only group level variance. It is tempting to compute the analogue of a multiple correlation coefficient to indicate how much variance is actually explained at each level (cf. Bryk and Raudenbush, 1992). However, this `multiple correlation' is at best an approximation, and it is quite possible for it to become smaller when we add explanatory variables (something that cannot happen with a real multiple correlation.) For a discussion of the problems and more sophisticated procedures see Snijders and Bosker (1994).

If we use an exploratory procedure to arrive at a `good' model, there is of course always the possibility that some decisions that have led to this model are based on chance. If the sample is large enough, we may split it in two, use one half for the model search, and the other for cross-validation. See Camstra and

Boomsma (1992) for a review of cross-validation procedures.

2.3 An Example of a Simple Two-Level Regression Model

The example below concerns the effect of interviewers and respondents on survey results. In survey research, the usual procedure is that there are a number of interviewers and that each interviewer questions many respondents. Thus, both interviewer and respondent characteristics can have an effect on the survey results, and much methodological research has been spent on the question how much interviewer and respondent bias is present in social survey data. Since respondents are nested within interviewers, in methodological terms this is clearly a multilevel problem. The specific example investigates how much interviewer and respondent characteristics influence the *speed* of interviewing (i.e., how many questions have been asked and answered in a given time period).

The data are from 515 respondents, interviewed by 20 interviewers. The dependent variable is the speed of interviewing, measured by the number of questions answered per minute. Leaving out nonsignificant variables, we have three explanatory variables at the respondent level: `tel' (whether the respondent was interviewed by telephone instead of face-to-face), `age' (the respondents age in years), and `lonely' (loneliness measured by a multi-item scale developed by de Jong-Gierveld, 1985). There are four explanatory variables at the interviewer level: `training' (amount of previous interviewer training), `pref.tel' (interviewer prefers telephone to face-to-face method), `extroversion' (interviewer extroversion), and `soc.ass.' (interviewer social assurance). There is one significant cross level interaction: the interaction between the variables `tel' and `soc.ass.' The results are reported in Table 2.1. below:

Table 2.1 Multilevel regression results interviewer/respondent data[a]

Fixed Part:	Regression coefficients:	p-value
Respondent level		
intercept	1.43	
tel	.30	.00
age	-.01	.00

lonely	-.04	.00
Interviewer level		
training	.25	.01
pref.tel	.27	.00
extro	.02	.00
soc.ass	.01	.15
Interaction		
tel × soc.ass	.01	.05
Random Part:	Variance components:	p-value
σ^2	.52	
$\sigma^2_{\text{intercept}}$	.03	.00
σ^2_{tel}	.01	.00
Deviance:	1155	

[a] *p*-values based on approximate standard errors provided by VARCL.

The interpretation of Table 2.1 is much like that of an ordinary multiple regression model. Interviews are faster with a telephone interview, and with younger and less lonely respondents. Trained interviewers, interviewers with preference for using the telephone, and extroverted interviewers are also longer. The effect of the interviewers' social assurance should be interpreted together with the effect of the interaction; careful inspection of the regression equation leads to the conclusion that in the face-to-face interview socially sure interviewers take more time for the interview.

2.4 Standardizing Regression Coefficients

The coefficients in Table 2.1 are unstandardized regression coefficients. To interpret them properly, we should take the scale of the explanatory variables into account. In multiple regression analysis (and structural models, for that matter) coefficients are often standardized because that facilitates the interpretation when one wants to compare the effects of different coefficients within one sample. (If the goal of the analysis is to compare different samples to each other, one should always use unstandardized coefficients!). To standardize the regression coefficients in Table 2.1, one could standardize all variables before putting them into the multilevel analysis. It is also possible to derive the standardized regression coefficients from the unstandardized coefficients:

$$\text{standard coefficient} = \frac{\text{(unstand. coeff.)} \times \text{(st.dev. explanatory var.)}}{\text{st. dev. dependent var.}} \qquad (2.13)$$

The variables in the interviewer example have quite different scales. For example: using telephone instead if face-to-face methods is indicated by a 0-1 dummy variable (0=telephone, 1=face-to-face) with a standard deviation of .50, while age is measured in years with a standard deviation of 17.71. If we apply formula (2.13) to the coefficients in Table 2.1 we get the standardized results in Table 2.2:

If we inspect the standardized regression coefficiens in Table 2.1, `extroversion' becomes the most important explanatory variable. Also `age' and `loneliness,' which seem negligeable in their unstandardized form in Table 2.1, now look fairly important. The reason is that these explanatory variables have a larger scale range. When the set of explanatory variables contains variables of widely different scales, reporting the standardized coefficients in addition to the raw coefficients helps the interpretation. In conventional multiple regression, this is practically standard practice. In multilevel regression, where the standard software does not automatically provide standardized coefficents, they are usually not reported. If we use equation (2.13) to calculate standardized coefficients, we should realize that the variance components still refer to the unstandardized coefficients. If we

need variance components for standardized coefficients, we must use the standardized variables in the analysis.

Table 2.2 Interviewer/respondent standardized coefficients

Fixed Part:	standardized regression coefficients:	p-value
Respondent level		
tel	.33	.00
age	-.22	.00
lonely	-.12	.00
Interviewer level		
training	.16	.01
pref.tel	.28	.00
extro	.36	.00
soc.ass	.13	.15
Interaction		
tel × soc.ass	.06	.05

2.5 Interpreting Interactions

Whenever there are interactions in a multiple regression analysis (whether these are cross-level interactions in a multilevel regression analysis or interactions in an ordinary regression analysis does not matter) there are two important technical points to be made. Both stem from the methodological principle that in the presence of a significant interaction the effect of the interaction variable and the direct effects of the explanatory variables that make up that interaction must be interpreted together as a system (Jaccard, Turrisi & Wan, 1990, Aiken & West, 1992).

The first point is that if the interaction is significant, it is best to include both direct effects in the regression too, even if they are not significant.

The second point is that in a model with an interaction effect, the regression coefficients of the simple variables carry a different meaning than in a model

without this interaction effect. If there is an interaction, then the regression coefficient of one of the direct variables is an estimate of the regression in the case that the other variable that is involved in the interaction is equal to zero, and vice versa. As a result, if for one direct variable the value `zero' is widely beyond the range of values that have been observed (as in age varying from 18-55), or if the value `zero' is in fact impossible for that variable (as in the social assurance scale where scores have a possible range from 14-98), the regression coefficient for the other variable has no substantive interpretation. In many such cases, the regression coefficient for at least one of the variables making up the interaction will be very different from the corresponding coefficient in the model without interaction. *This change does not mean anything.* One remedy is to take care that the value `zero' is meaningful and actually occurs in the data. One can accomplish this by centering both explanatory variables around their overall mean.[1] After centering, the value `zero' refers to the mean of the centered variable; in this case the regression coefficients do not change when the interaction is added to the model. When the explanatory variables are centered, the regression coefficient of one of the variables in an interaction can be interpreted as the regression coefficient for individuals with an `average' score on the other variable. If all explanatory variables are centered, the intercept is equal to the grand mean of the dependent variable.

In practice, to interpret an interaction, it is helpful to write out the regression equation for one explanatory variable for various values of the other explanatory variable. When both explanatory variables are continous, we write out the regression equation for the lower level explanatory variable, for a choice of values for the explanatory variable at the higher level. Possible choices are the mean, maximum and minimum, or the median and the 25th and 75th percentile. A plot of the regression lines generally clarifies the meaning of the interaction.

In the interviewing example there is one cross level interaction in the model between `telephone condition' and `social assurance.' The telephone condition-

[1]Standardizing the explanatory variables has the same effect. In that case I recommend not to standardize the interaction because that makes it difficult to compute predictions or plot interactions. Standardized regression weights for the interaction term can always be determined with equation (2.13).

variable is scored 0=face-to-face and 1=telephone: the value `zero' clearly has an empirical interpretation here. Centering is not needed here. Since in this example centering would refer to the `average method used', an ambiguous concept at best, centering is not attractive either. Thus, the regression coefficient b=.01 (in Table 2.1) for `social assurance' refers to the face-to-face situation. In the telephone situation this regression coefficient becomes .01+1*.01=.02 (the regression coefficient plus the number implied by the interaction when tel=1). Social assurance is measured on a 14-item scale with items scored 1-7; theoretically the scores could vary from 14 to 98, but in fact the range in the sample is from 41 to 78 with a mean of 61.76. In this case centering is attractive; if we use centering the value of the regression coefficient for the telephone condition refers to the situation where we use interviewers of average social assurance. In Table 2.1 `social assurance' was in fact centered around its mean value.

Another way to make interactions easier to interpret is to plot the regression slopes for one of the explanatory variables for some values of the other. In this case, since `tel' has only two values, we plot the regression slope of `social assurance' for tel=0 and for tel=1. Figure 2.1 below shows this interaction; for interviewers with a low social assurance there is almost no difference in speed between the face-to-face and the telephone condition. The more social assurance the interviewers have, the longer the interview takes, but only in the face-to-face condition.

3. Working with HLM, VARCL and MLn

The three programs that are currently the most popular programs for analyzing the multilevel regression model are HLM (Bryk, Raudenbush, & Congdon, 1994), VARCL (Longford, 1990), and ML3/MLn (Prosser, Rasbash & Goldstein, 1991; Rasbash & Woodhouse, 1995). To highlight both the similarities and the differences between these programs, and to provide some feeling for how multilevel analyses may proceed, this chapter provides a systematic analysis of a small data set with the programs HLM, VARCL, and MLn. In theory, each program should lead to the same conclusion. In practice, this need not be the case. There are small differences in the estimates produced by the programs (cf. Kreft et al., 1990) that might turn out to be important in fitting models to real data. The programs and their users' guide all use a different notation. For the sake of consistency, I use the notation introduced in the previous chapter, and only briefly discuss the differences between this notation and that of the specific program being discussed. The programs also differ in the layout and the amount of information given in the output. Again, I will display the results in a standard format, and briefly point out important differences with the computer output. Finally, the programs differ sharply in the extra features offered to the analyst for exploring the data set. In the analysis of our example data, I have *not* tried to follow exactly the same procedure (for instance the exploratory procedure outlined in chapter 2) with each program. Instead, I have tried to follow a procedure that feels `natural' given the specific program used and its particular features. This approach has the added advantage that it provides some sense of the differences between the programs in `look and feel.'

The example data were collected by Van der Wolf (cf. Hox & de Leeuw, 1986). In this study, 681 pupils from 29 classes were questioned at the beginning of the school year (pretest), and again at the end of the school year (post test). After removal of all cases with missing values, we have a data set which consists of 428 pupils in 28 classes. The dependent variable is `pupil loneliness', measured at the end of the school year. There are four explanatory variables at the pupil level, all measured at the beginning of the school year: X_1=pupil gender, X_2=repeat (how

many times a pupil had to repeat a class), X_3=ethnicity (parent not Dutch), and X_4=loneliness pretest. There are three class level variables, all three teacher characteristics: Z_1=teacher gender, Z_2=teacher experience, and Z_3=teacher having taken extra postgraduate courses. In the remainder of this chapter, the equations use the X's and Z's, and the text uses the full variable names.

Since HLM is the easiest program to use, and MLn the most difficult, with VARCL lying in between, the first analysis uses HLM, the second VARCL, and the last MLn.

3.1 HLM Analysis of the Example Data[1]

HLM (Bryk, Raudenbush & Congdon, 1994) is a set of related programs for 2-level and 3-level analysis, called HLM/2L and HLM/3L respectively, and a special program called VKHLM that can be used for meta-analysis (cf. section 4.1). The HLM output contains the parameter estimates, their standard errors, the (co)variances at the two levels, and the deviance. For most of the parameter estimates HLM provides *p*-values as an indicator for their significance. Furthermore, the variance in the β_j coefficients is partitioned into sampling variance and true residual parameter variance, the latter can in principle be explained by second level variables. HLM produces estimates of the `parameter reliability', which is the proportion of true parameter variance in each parameter, and the corresponding *p*-value for the null-hypothesis that the true parameter variance of a specific β_j is zero. As noted in section 2.2, HLM has a special approach to estimating the *p*-values, which may make a noticeable difference when the number of groups is small.

Our example data have two levels, with explanatory variables at both levels. Chapter Two presents the basic two-level regression model for one explanatory variable at each level (equation 2.1). The model was written as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + e_{ij} \qquad (3.1)$$

[1]The analysis reported here uses HLM/L2 for two levels. HLM/L3 for three levels has more options, especially to test the assumptions of the model. See Bryk et al., 1994, for details.

for the first (lowest) level, and the intercept β_{0j} and the regression slope β_{1j} were written as linear regression functions of the second (higher) level explanatory variable Z:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j}, \tag{3.2}$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \tag{3.3}$$

Substitution of (3.2) and (3.3) into (3.1) gives the one-equation version of the model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij} \tag{3.4}$$

The standard HLM notation differs from this notation. The notation used in the HLM manual (Bryk et al., 1994) closely follows the notation in Bryk and Raudenbush (1992). The equations take the form:

$$Y_{ij} = \beta_{0j} + \beta_{1j} X_{ij} + r_{ij} \tag{3.5}$$

for the individual level, which HLM calls the *unit* level, and

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + U_{0j}, \tag{3.6a}$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + U_{1j}. \tag{3.6b}$$

for the group level.[1] The intercept β_{j0} is called the *base* in HLM's terminology. The random error r_{ij}, which is our e_{ij} in equation 3.1, is by default assumed to have the same variance in all groups, an assumption which in HLM can be relaxed. The variances and covariances of the beta's are in the covariance matrix T (Tau), which is Σ (sigma) in our notation of chapter two. HLM produces rather voluminous output, including the starting values for the computational procedure. (Sometimes

[1]The earlier version of HLM (Bryk et al., 1988) used a slightly different notation in the manual, with the symbol theta instead of gamma in the second level regression equation.

these values may be of interest in their own right, but generally they can be ignored.)

As mentioned in the introduction of this chapter, I will use the notation introduced in chapter two, rather than the different notations used by each program.

In the general multilevel regression model for our example data, the intercept and the regression coefficients of the four pupil level variables vary at the class level. Thus, the pupil level model is:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + e_{ij}. \qquad (3.7)$$

All regression coefficients β are assumed random, and the class level model which models them using the teacher variables Z is:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{03}Z_{3j} + u_{0j} \qquad (3.8a)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11}Z_{1j} + \gamma_{12}Z_{2j} + \gamma_{13}Z_{3j} + u_{1j} \qquad (3.8b)$$
$$\beta_{2j} = \gamma_{20} + \gamma_{21}Z_{1j} + \gamma_{22}Z_{2j} + \gamma_{23}Z_{3j} + u_{2j} \qquad (3.8c)$$
$$\beta_{3j} = \gamma_{30} + \gamma_{31}Z_{1j} + \gamma_{32}Z_{2j} + \gamma_{33}Z_{3j} + u_{3j} \qquad (3.8d)$$
$$\beta_{4j} = \gamma_{40} + \gamma_{41}Z_{1j} + \gamma_{42}Z_{2j} + \gamma_{43}Z_{3j} + u_{4j} \qquad (3.8e)$$

HLM follows the two-equation notation of (3.5) and (3.6). To start the HLM analysis, we must prepare two data files: one containing the pupil level data, and one containing the class level data. Both files must be sorted on a class identification variable, which HLM uses to link the two data files. When the raw data have been read in and the variables have been named, HLM asks whether the variable names and sufficient statistics should be saved in a so-called `sufficient statistics file'. Since this makes subsequent runs faster, creating a sufficient statistics file is a good first step in an HLM analysis. HLM also requires a default file which sets some program defaults. Bryk et al. (1988) suggest speeding up the program by setting the maximum number of iterations to 10 for exploratory analyses, and to a much higher maximum for the final model. As Kreft et al. (1990) show, a maximum of 10 iterations is often too small for programs that, like HLM, use the EM algorithm. In our analyses, we have set the maximum

number of iterations at 10, as recommended by Bryk et al (1988). When convergence is not reached after 10 iterations, the model is examined, and if it looks interesting, follow up analyses are done with the maximum number of iterations set much higher (e.g., at 100).

After reading in the data, HLM prompts the researchers with the options:

> Do you wish to:
> Examine means, variances, chi-squared, etc?
> Specify an HLM model?
> Define a new outcome variable?
> Exit?

The option `examine means, variances, chi-squared, etc.?' is useful if researchers want to explore the data before actually fitting a hierarchical model. Choosing this option provides us with the following information:

Table 3.1 Preliminary HLM analysis, outcome variable Lonely Posttest

Potential independent variable:	Mean univariate regression coefficient:	ANOVA estimate of variance in regression coefficient	Reliability	Chi-square	K
Means	-.00065	.32686	.86263	181.0216	28
Pup.gender	-.33575	.12143	.37488	42.80946	27
Repeat	.32611	est < 0		18.20512	26
Ethnic	.39092	est < 0		23.71305	28
Lonely pre	.69961	.00190	.02370	37.03086	28

The information from the preliminary analysis in Table 3.1 is based on simple univariate analyses of variance and single level regressions. Since these are only a rough indication of what we may expect to find in a subsequent multilevel analysis, we should not look for fine distinctions here. The chi-square tests give no p-value, which, given the approximate nature of these single level tests, is probably just as well. To interpret these figures, it is useful to know that when the

null-hypothesis is true, the expected value of a chi-square variate is approximately equal to the number of degrees of freedom, which is K-1 (the number of groups minus one); the number of groups (K) is also given in Table 3.1. Using this fact, we see that there is clear evidence of significant between group variation in the intercepts (HLM calls these the means: the relevant chi-square is 181.0 with 28 degrees of freedom), and some evidence of significant between group variation in the slope coefficient for pupil gender (chi-square is 42.8 with 27 degrees of freedom).

The variation in the slope coefficient for the loneliness pretest is probably not significant. The column under `reliability' gives an estimate of the proportion true parameter variance (as opposed to sampling variance) in the random coefficients. The proportion of reliable variance in the means is estimated as 0.86, and in the slopes for pupil gender as 0.37. The proportion of reliable variance in the slopes for the loneliness pretest is estimated as 0.02, which is very small, although it could still turn out to be significant when a multilevel model is computed. The variance estimates for the regression coefficients of `repeat' and `ethnic' are negative, with chi-square values lower than the corresponding number of degrees of freedom; in this case the reliability estimate is not computed. The interpretation of these results is that these regression coefficients have zero variance. Consequently, when we continue our analysis and specify a multilevel regression model, we can specify these regression slopes as fixed, instead of random.

After this preliminary analysis, we specify a multilevel (hierarchical) regression model. The first model fitted is the `intercept only' model, which models the intercept as random, but has no explanatory variables at any level. The within class model is:

$$y_{ij} = \beta_{0j} + e_{ij}, \tag{3.9}$$

and the between class model is:

$$\beta_{0j} = \gamma_{00} + u_{0j}. \tag{3.10}$$

This model contains only one gamma (the intercept γ_{00}), and two variance components: the variance of u_{0j}, which is σ_{00} (the HLM output gives this as the

element base*base in the covariance matrix Tau, which in this case has only one element), and the variance of the e_{ij}, which is the common within group variance σ^2 (sigma squared). For the example data the intercept variance is 0.32, and sigma squared is 0.72. This implies that the between group variance of the intercept is (within rounding errors) 32% of the total variance. Another way to put this is that the intra class correlation is estimated as 0.32.

The first substantively interesting HLM model employs all four variables at the pupil level, but no variables at the teacher level. The within class model is:

$$y_{ij}=\beta_{0j} + \beta_{1j}X_{ij1} + \beta_{2j}X_{ij2} + \beta_{3j}X_{ij3} + \beta_{4j}X_{ij4} + e_{ij} \qquad (3.11)$$

By default HLM assumes that all regression coefficients in equation (3.11) have random variation. Thus, the between class model for all beta's is:

$$\beta_{pj} = \gamma_{p0} + u_{pj} \qquad (3.12)$$

We have a total of five gamma coefficients, one for the intercept of the within class model, and one for the slope of each of the four explanatory variables (X_1 to X_4) in the within class model. We also have five error terms u_{pj} at the class level and the usual error term e_{ij} at the individual pupil level.

Before the calculations start, HLM asks whether the explanatory variables should be centered around the class means. To keep the results comparable to the other programs, we have not used this option. If explanatory variables at the individual level are not explicitly declared as fixed, HLM considers them by default to be random. Thus, all regression coefficients in equation (3.11) are assumed by default to be random, which means that all error terms u_{pj} in equation (3.12) are assumed to be non-zero. For this model, HLM did not converge in 10 iterations. It did also not converge in 30 or 100 iterations.

As I mentioned earlier, if the numerical estimation procedure does not converge, this is often a sign that something is wrong with the model. A good advice in such cases is to change the model into a model with fewer random parameters. As a matter of fact, the non-converged (thus: inaccurate) multilevel estimates of the variance components for the regression coefficients and their

associated p-values in Table 3.2 below confirm the conclusion from the preliminary analyses in Table 3.1, that the slopes (regression coefficients) of the pupil level variables `repeat' and `ethnic' may be considered identical in all classes (fixed).

Table 3.2. HLM model with pupil variables only, results after 100 and between parentheses results after (10) iterations.

for:	gamma		p		var. of u		p	
intercept	.06	(.05)	.35	(.36)	.26	(.24)	.00	(.00)
gender	-.15	(-.15)	.07	(.06)	.07	(.03)	.09	(.10)
repeat	.16	(.15)	.08	(.12)	.03	(.05)	.44	(.46)
ethnic	.09	(.19)	.19	(.17)	.05	(.04)	.39	(.41)
lon. pre	.64	(.65)	.00	(.00)	.03	(.03)	.18	(.19)

While the differences between the results at 10 and at 100 iterations could be important when regression coefficients or variances of borderline significance are considered, in our case it is unlikely that the results after many more iterations would lead to a different conclusion about the two coefficients for `repeat' and `ethnic,' so no further computations were done. Based on the preliminary analyses in Table 3.1 and on the results in Table 3.2, I assume that the slope coefficients for the variables `repeat' and `ethnic' are fixed.

The next model, with fixed within class regression slopes for the pupil variables `repeat' and `ethnic,' also encounters convergence problems. Again the results at 10 and 100 iterations do not differ much. The variance for the slope of `loneliness pretest' is again estimated as 0.03, with a p-value of 0.17. Thus, in the next model this slope is also fixed. Now the model converges very fast; it needs only four iterations. The two remaining parameter variances of 0.17 for the intercept and 0.09 for the slope of `pupil gender' are both significant at $p<0.01$. The equations for this model are, at the pupil level:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + e_{ij} \qquad (3.13)$$

and at the class level:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (3.14a)$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \quad (3.14b)$$
$$\beta_{2j} = \gamma_{20} \quad (3.14c)$$
$$\beta_{3j} = \gamma_{30} \quad (3.14d)$$
$$\beta_{4j} = \gamma_{40} \quad (3.14e)$$

The results for this model are reported in Table 3.3 below:

Table 3.3. HLM model with pupil variables only

for:	gamma	p	var. of u	p
intercept	.06	.32	.17	.00
gender	-.17	.07	.09	.01
repeat	.14	.10		
ethnic	.12	.08		
lon. pre	.63	.00		

HLM computes a deviance for the model tested, based on the likelihood function. When one model is a subset of another model, the difference between their deviances is distributed as a chi-square, with degrees of freedom equal to the difference of the number of parameters included in the two models. As part of its output, HLM prints the deviance and the number of parameters in the model; these can be input in a subsequent model and HLM will then perform the appropriate chi-square test. To use this test properly, it is important to realize that the two-level version HLM/2L computes a restricted maximum likelihood solution (RML, see chapter 2). This means that the regression coefficients (intercept and slopes) do not enter the likelihood function. As a result, we cannot use the difference between deviances to test the difference between two models that differ *only* in their regression coefficients. When we compare two models that only differ in their regression coefficients, we will see that the number of parameters estimated, as reported by the program, is the same, even if the actual value of the deviance may be somewhat different. As a result, in HLM the chi-square test based on the difference between the deviances of two nested models can only be used to test differences in the random parts (the variance components) of those models. For the regression coefficients we have to rely on the standard errors of the regression coefficients and their associated p-values, which are computed for

each regression coefficient separately. However, these tests are only approximate and should therefore be interpreted with some caution. Founding important decisions based on a borderline (non)significance of one of these coefficients is risky. The three-level program HLM/3L uses Full Maximum Likelihood, which makes it possible to use the likelihood test for regression coefficients as well.

All the changes we have so far made in the model are changes in the random part. As a consequence, we can compare the models by looking at their deviances and at the number of parameters estimated in the random part of the model. For the first model (results in Table 3.2), HLM reports 16 parameters estimated and a deviance of 894.2. The next model, with only the within class regression slopes for `repeat' and `ethnic' fixed, estimates seven parameters, and it has a deviance of 900.8. The difference between the two deviances is 6.6, which is approximately distributed as a chi-square variate with 16-7=9 degrees of freedom.[1] The p-value of this chi-square is p=0.68, which is not significant. The final model (results in Table 3.3) estimates four parameters, and has a deviance of 908.3. This is not significantly different from the previous model (χ^2=7.5, df=5, p=0.19), or from the first model (χ^2=14.1, df=12, p=0.29). From the chi-square tests of the difference between deviances of successive models, we conclude that omitting these (co)variances from the random part does not significantly affect the overall model fit.

In the last model, with the intercept and the slope of the variable `pupil gender' random (all other coefficients are fixed), the pupil level error variance sigma squared is 0.39, which is considerably smaller than the sigma squared of the intercept-only model, which is 0.72. We might want to compute another intra class correlation, or want to determine how much intercept variance is explained. However, this would be misleading, since we now have not only a random intercept, but also one random slope for `pupil gender.' The residual error terms $u_{interc.}$ and $u_{gender}$ are correlated, and the variance of $u_{interc}$ $\sigma_{00}$ depends on the way the explanatory variable `pupil gender' is scaled. In these data, gender is coded 1=boy, 2=girl. If we change this code to boy=0, girl=1, or to boy=-1, girl=+1 (all

[1]Since the first two models did not converge, the deviances are imprecise. Still, close to the maximum of 100 iterations, the deviance did not change much between iterations, and therefore I decide to use them nevertheless.

permissible transformations for interval data), we will get slightly different variance estimates. Thus, the specific value of this `intra class correlation' would vary with (quite permissible) linear transformations of `pupil gender.' The conclusion is that, when we introduce random slopes, interpretations of variance components should be made with caution.

After each analysis, HLM asks whether the group level variables specified before should be regressed on the `residuals' from the within group analysis. This means that we may now attempt to explain the different values across classes for the random regression coefficients (both the intercept and the slope of `pupil gender') by the teacher level variables Z_1 to Z_3 (teacher gender, teacher experience, and teacher having followed courses), using simple regression techniques. This makes sense only for those regression parameters that have large (significant) random variation. In Table 3.2, ethnicity and repeat showed almost no variation over classes. Trying to explain this variation by teacher variables at the class level does not make sense, since there is no reliable variation in these slope coefficients (as is also shown in Table 3.1). Trying to explain the significant variation over classes in the intercept and in the slope (cf. Table 3.3) for pupil gender does make sense.

Since ordinary single level regression estimates can be computed much faster than the corresponding multilevel estimates, this option gives us an opportunity for a quick `preview' of what may happen if we include specific second level explanatory variables to model the random coefficients. If we choose this option, HLM reports ordinary regression slopes, standard errors, and T's (which are normal deviates), but no p-values. For the intercept and the slope of `pupil gender', the two random parameters in the final model (cf. Table 3.3), HLM reports only one effect with a T-value exceeding ± 2.0: the interaction between `pupil gender' and `teacher experience'. The effects of `teacher gender' and `teacher experience' on the intercept have T-values exceeding ± 1.0. Since these results are from an ordinary analysis, the p-values are unreliable and biased in the direction of producing too many `significances'. If the number of explanatory variables was large, it would make sense to try further HLM analyses with only the one interaction exceeding T=2.0. In this case, since the number of class level explanatory variables is small, we may as well try all three effects that had a T

exceeding ± 1: the two direct effects of the teacher variables `teacher gender' and `teacher experience', and the interaction effect of `teacher experience' and `pupil gender'. The results are presented in Table 3.4, which is set up in a format similar to the way the HLM output is organized:

Table 3.4. HLM model with pupil and class variables

for:	gamma	SE	p	var. of u	p
intercept	.11	.34	.21	.17	.00
*teach. gndr	.30	.17	.09		
*teach. exp.	-.02	.02	.23		
pup. gender	.15	.19	.29	.07	.03
*teach. exp.	-.03	.01	.08		
repeat	.15	.09	.09		
ethnic	.12	.07	.08		
lonely pretest	.64	.04	.00		

Table 3.4 corresponds to the following model for the pupil level:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + e_{ij} \qquad (3.15)$$

and the class level model is:

$$\beta_{0j} = \gamma_{00} + \beta_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} \qquad (3.16)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{12}Z_{2j} + u_{1j} \qquad (3.17)$$

The matrix of variances and covariances of u_{0j} and u_{1j} at the class level is (the HLM output gives this as the matrix Tau):

$$\Sigma = \begin{matrix} \sigma_{00} & \sigma_{10} \\ \sigma_{01} & \sigma_{11} \end{matrix}$$

Equation (3.17) makes clear that the entry pup. gender * teach. exp. is the cross-level interaction between the pupil level variable `pupil gender' and the class level variable `teacher experience'.

Generally, underspecification of models is considered a larger danger than overspecification (cf. Mosteller & Tukey, 1977). To avoid underspecification, I

decide to continue the analysis by omitting only those effects that are not significant at a significance level of 0.10. We then finally arrive at the results in Table 3.5:

Table 3.5 Final HLM model with pupil and class variables

for:	gamma	SE	p	var.of u	p
intercept	.06	.10	.32	.17	.00
pup. gendr	.19	.18	.22	.07	.02
*t. exp.	-.03	.01	.04		
repeat	.15	.09	.09		
ethnic	.12	.07	.09		
lonely pre	.63	.04	.00		

As I noted in chapter two, predicting a slope coefficient by higher level explanatory variables in a higher level regression equation is statistically and conceptually equivalent to introducing a cross-level interaction term between a class level variable and a pupil level variable. HLM obscures such interaction terms by using equations 3.15 to 3.17 to describe the hierarchical regression model. The model implied by the results in Table 3.5 corresponds to the following single model equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{12}X_{1ij}Z_{2j} + \\ + u_{0j} + u_{1j}X_{1ij} + e_{ij} \qquad (3.18)$$

This is important, because interaction terms tend to correlate highly with the explanatory variables which make up the interaction. Therefore, to interpret any interaction correctly, the direct effects of these explanatory variables must also be included in the regression equation (Jaccard et al., 1990). The single equation version in equation (3.18) makes clear that by successively omitting nonsignificant terms, we have ended up with a regression equation that does not follow this rule. Since the effect of `teacher experience' on the slope of `pupil gender' is an interaction, both `teacher experience' and `pupil gender' must be in the regression model too. Thus, the model in equation (3.18) should be replaced by a new model that includes all necessary terms. This leads to the model in equation (3.19):

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{02}Z_{2j} + \gamma_{12}X_{1ij}Z_{2j} + \\ + u_{0j} + u_{1j}X_{1ij} + e_{ij} \qquad (3.19)$$

To clarify what happens, I have rearranged the next table to present first the main effects at the pupil and the class level, followed by the cross-level interaction between pupil gender and teacher experience:

Table 3.6 Corrected final HLM model with pupil and class variables

for:	gamma	SE	p	var. of u	p
intercept	.17	.22	.28	.18	.00
pup. gend.	.15	.19	.28	.07	.00
repeat	.15	.09	.10		
ethnic	.12	.07	.08		
lonely pre	.63	.04	.00		
teach. exp	-.01	.02	.32		
pup. gend. * t. exp.	-.03	.01	.08		

Table 3.6 makes clear that there is a cross-level interaction between X_1*Z_2 in the model, a fact that remains hidden in the way HLM organizes its output. Substituting the higher level regression equations into the lower level regression equation, as at the start of this section, shows this relation between the explanatory variables more clearly. For the same reason, a format as in Table 3.6 is to be preferred to the format used in Table 3.4 and 3.5, which follows the HLM output; Table 3.6 makes the presence of the interaction more evident.

The conclusion from all these analyses is that there is significant between class variation in the intercept and in the slope for pupil gender. The class level variables at our disposal cannot explain the class level variation of the intercepts. The variation in the regression slope for pupil gender can partly be explained by the teacher's experience, but there remains significant unexplained parameter variance. The loneliness pretest is significant in all models, but this information is not very interesting, because this is simply the pretest, which is a covariate. I

interpret the p-values of 0.10 and 0.08 for `repeat' and `ethnic' as a sign that these variables merit further research with a larger sample size. It is interesting that there is no significant overall effect of pupil gender on the dependent variable `loneliness,' while this effect does vary significantly across classes. The fact that it shows significant variation tells us that this apparent `nonsignificance' masks an interaction with class level explanatory variables. The cross-level interaction between pupil gender and teacher experience tells us that classes with an experienced teacher have a different effect of pupil gender on loneliness than classes with inexperienced teachers.

In interpreting the interaction in the interviewer example in chapter two I have noted that an interaction between two explanatory variables X_1 and X_2 means that the regression coefficient for X_1 has a different value for each value of X_2 (cf. Jaccard et al., 1990). The value given for the regression coefficient of X_1 is the value for $X_2=0$, and the value given for the regression coefficient of X_2 is the value for $X_1=0$. If X_1 and/or X_2 do not (or sometimes even cannot) attain the value `zero,' the values of the regression coefficients of $X_1$ and $X_2$ can be quite misleading if they are interpreted on their own. In our example, `pupil gender' is scored 1 for boys, 2 for girls, and `teacher experience' is measured simply in years. Thus, `pupil gender' cannot become zero, and teacher experience is very unlikely to be zero. This means that as soon as the interaction is included in the model, the regression coefficients for pupil gender and teacher experience have no longer a simple interpretation, because they refer to a situation that cannot exist in our data. In fact, if we compare the estimate for the regression coefficient of pupil gender in a model without the interaction (Table 3.3) with the estimates in the model that includes the interaction (Table 3.6), we see that it changes from a negative -.17 to a positive +.15. So, in interpreting the final model, we would be tempted to conclude that girls are more lonely than boys, but this conclusion would only hold for classes that have teachers with zero experience. If we ignore the teacher variables, (Table 3.3) we find that girls are actually less lonely than boys.

To interpret this interaction, it is helpful to write out the regression equation for one explanatory variable for various values of the other. With continuous variables, a good strategy is to write the regression for X_1 for the mean and for the minimum and maximum of X_2 (the median and the 25th and 75th percentile

would also be a good choice). The pupil level variable `pupil gender' is scored 1 for boys, 2 for girls, and the class level variable `teacher experience' is measured in years. Since the pupil variable is the one with the varying slopes, we would generally write the regression equation for the pupil variable for different values of the teacher variable. Since in this specific case the pupil variable `gender' has only two values, while `teacher experience' has many, I prefer to write the regression equation for teacher experience for the two values of pupil gender. Ignoring the intercept and the other explanatory variables, the regression equations relating posttest loneliness to teacher experience for boys and girls separately can be found by taking the appropriate part of the regression equation (the relevant numbers are in Table 3.6), filling in the values for pupil gender and teacher experience, and working out the equation. The resulting equation is, for boys: Y=0.15-0.04*teach.exp., and for girls: Y=0.30-0.07*teach.exp. If I work out the predicted loneliness for girls and boys, for different values of teacher experience, I find that for teachers with less than five years of experience, girls are more lonely than boys, and for teachers with six or more years of experience, boys are more lonely than girls. (In the interview example in chapter two, I visualized the interaction by plotting both regression lines in one figure, which simplifies interpretation). Since in our data set the average teacher experience is more than 12 years, much larger than the turnover point of five years, on the average boys are more lonely than girls, just the opposite of what one might expect from the positive value of 0.15 for the gamma (main effect) of pupil gender in the last model (Table 3.6). This is analogous to what happens if we interpret a main effect in an analysis of variance in the presence of a strong interaction. The conclusion is that, just as in analysis of variance, it is dangerous to interpret direct effects in the presence of interactions.

It is important to note again that all these complicated calculations and interpretations can be avoided by scaling the explanatory variables in such a way that `zero' is an interpretable value that also is observed in the data. With pupil gender we can accomplish this by scoring boys=0, girls=1 instead of the boys=1, girls=2 scoring now used. A more general strategy is to *center* the explanatory variables around their overall mean. With centering, regression slopes do not change if an interaction is added to the model, and the size of the direct effect of a

variable involved in an interaction can be interpreted as the effect of that variable for pupils with an average value on the other variable.[1]

Some coefficients reported in Table 3.6 are of borderline (non)significance. Since the significance levels are approximate we may decide to keep these in the final model, and decide to replicate the study with a larger sample.

3.2 VARCL Analysis of the Example Data

I repeat the single-equation version of the basic two-level regression model:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij}$$

The general two level model in VARCL notation (cf. Longford, 1990) is:

$$Y_{ij} = b_{j1}X_{ij1} + b_{j2}X_{ij2} + b_{j3}X_{ij3} + \ldots + b_{jk}X_{ijk} + e_{ij} \qquad (3.20)$$

where the subscript i refers to the first level, j to the second level, and k to the regression coefficients. The first coefficient b_{j1} is the intercept; the program automatically sets the associated X_{ij1} to one. As a consequence, in VARCL the variable X_{ij1} is always the constant 1.0, which makes b_{j1} the regression intercept. The errors e_{ij} are assumed to have a normal distribution with mean zero and a variance σ^2. At the second level the random coefficients b_{jk} are written as:

$$b_{jk} = ß_{jk} + \delta_{jk} \qquad (3.21)$$

where the $ß_{jk}$ are regression parameters to be estimated, and the δ_{jk} are residual error terms. The residual errors δ_{jk} are assumed to be normally distributed with a mean of zero and covariance matrix Σ. The difference of equation (3.21) with our notation is that I use γ (gamma) for the regression coefficients in equation (3.21), and u_{jk} for the second level error. Thus, compared to equation (2.1) to (2.4), VARCL

[1]Note that the centering option offered in HLM centers the variables around their group means, and not around the overall mean. Centering around the group mean has a totally different effect on the interpretation of the regression weights.

estimates models of the form:

$$Y_{ij} = ß_{0j} + ß_{1j}X_{ij} + e_{ij} \qquad (3.22)$$

for the first level. The random intercept $ß_{0j}$ is written as the usual linear function of the second level explanatory variable Z:

$$ß_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \qquad (3.23a)$$

The random slope $ß_{1j}$, however, is written as:

$$ß_{1j} = \gamma_{10} + u_{1j}. \qquad (3.23b)$$

Substituting (3.23a) and (3.23b) into (3.22) gives (3.24):

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + u_{1j}X_{ij} + u_{0j} + e_{ij}. \qquad (3.24)$$

If we compare my equation (2.4) to the equivalent VARCL equation given in (3.24), we see that VARCL contains no build-in provision for modeling lower level regression slopes by higher level variables. To model the lower level regression slopes, we must compute the necessary interaction variables outside VARCL, and include these in the data set with the other variables. To explain how this works (see also my discussion of the random coefficient model in chapter 2), I go back to the example with one pupil level variable X_{ij} and one class level variable Z_j. The pupil level model is given by equation (3.22), which is repeated here:

$$Y_{ij} = ß_{0j} + ß_{1j}X_{ij} + e_{ij} \qquad (3.22)$$

The random coefficients $ß_{0j}$ (the intercept) and $ß_{1j}$ (the slope for X) are modeled by VARCL as: $ß_{0j}=\gamma_{00}+\gamma_{01}Z_j+u_{0j}$ (equation 3.23a) and $ß_{1j}=\gamma_{10}+u_{1j}$ (equation 3.23b). By substitution into (3.22) we get (3.24), which is also repeated here:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}Z_j + u_{1j}X_{ij} + u_{0j} + e_{ij}. \qquad (3.24)$$

Equation (3.24) is in fact the model we automatically get when we specify in the

program that Z is a class level variable, X is an individual level variable, and that the regression coefficient for X is to be random. Equation (3.20) is very similar to our equation (1.4), but it lacks the interaction. To model the random coefficient $ß_{1j}$ with the second level variable Z, we must expand (3.23b) by writing: $ß_{1j}=\gamma_{10}+\gamma_{11}Z_j+u_{1j}$. If we substitute that into (3.22) we get equation (2.4):

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + \\ + u_{1j} X_{ij} + u_{0j} + e_{ij} \qquad (2.4)$$

Comparing equation (2.4), which we want to estimate, with equation (3.24), which is normally estimated by VARCL, it is easy to see that we can fit (2.4) in VARCL by providing the program with the interaction-variable (Z_jX_{ij}) and entering it in the regression equation at the pupil level. Because the random part in (3.24) is already equal to that in (2.4), the regression coefficient for the interaction variable must be declared to VARCL as fixed; if it was declared as random VARCL would put an unwanted extra variance component in the model.

To analyze a two level model with VARCL, we must prepare three input files: a basic information file and two data files, one for each level (a three level model needs the basic information file plus three data files). The basic information file provides the program with the number and names of the variables at each level, the number of observations at each level, and some other information. The data in the two data files have to be sorted. No unit identifications are needed, because the basic information file specifies how many observations are in each group at each level. VARCL requires that the analyst defines a so-called `maximal model', which contains all variables and parameters that are thought to be theoretically interesting. VARCL allows for fitting many models in a single computer session, but all these models must be sub-models of the maximal model that is declared at the beginning of the analysis session.

A sub-model is a model that can be derived from the maximal model by constraining parameters from random to fixed, or dropping explanatory variables. If one model is a subset of another model, the difference between their deviances is approximately distributed as a chi-square with degrees of freedom equal to the difference of the number of parameters estimated in both models. Here, VARCL

differs from HLM in a very important aspect. VARCL uses full maximum likelihood (FML) estimation, where the fixed regression parameters are included in the likelihood function. When the number of parameters in a model is counted, the fixed part of the regression parameters is included, and as a result the test based on the difference in two deviances can also be used to test models that differ in the number of explanatory variables (regressors), but do not differ in their random parts. When Restricted Maximum Likelihood (RML in HLM/2L, RIGLS in MLn) is applied, only the deviances of models that differ in their random parts can be compared to each other. This means that to decide which sub-models merit further examination, analysts using VARCL can choose to compare the *deviance* of different models, or to inspect the *estimates* and *standard errors* of various coefficients in one specific model. Generally, when a number of variables is dropped from a model because their regression slopes are not significant, the chi-square test on the difference in deviances between the larger and the smaller model is also not significant. If both tests lead to different conclusions (which may happen, especially with models for non-normal data), the procedure using the deviance is considered to be more precise (McCullagh & Nelder, 1989).

VARCL provides parameter estimates and standard errors, but no p-values. If researchers want these, they have to be computed outside the program. Since the standard errors are only approximate, it is much easier to consider a parameter `significant' when its absolute value is at least twice as large as its standard error; this is approximately equal to a two-sided test at the five percent significance level.

After the first model has been calculated, VARCL offers the option to save the sufficient statistics and other information for the maximal model in a so-called dump file, which can be read directly by the program in later runs. Using a dump file enables the researcher to save time later by bypassing the model definition stage.

In addition to the maximal model, which is generally the most complex model that we are willing to consider, the `intercept only' model (a model that includes no explanatory variables) serves as a minimal model. The intercept only model is given by:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij} \qquad (3.25)$$

This model produces a value for the deviance, and estimates the pupil level variance σ^2 and the class level variance σ_{00} with its standard error.[1]

For the example data, the intercept only model has a deviance of 1128.3; it estimates the class level variance as 30 percent of the total variance, and the pupil level variance as 70 percent of the total variance.

Since the number of explanatory variables in our example is small, we can include all variables in the maximal model. To fit interaction terms with the program, we have to include all interaction variables in the data file. This forces the researcher to think of interactions that may be useful to include in the models to be tested, and to compute these interactions before executing VARCL. Our four pupil and three teacher variables define (4×3=) 12 interactions between the two levels. This number is not prohibitively large, but clearly in a larger data set the number of possible interactions can increase rapidly. Therefore, I start the VARCL analysis with a maximal model that contains no interactions, only additive effects, and attempt to determine at a later stage which interactions must be selected for further testing. To decide which model to keep, I will mainly focus on the deviances of the different models. The difference in the deviance of two nested VARCL models is chi-square distributed, with as degrees of freedom the difference in the number of parameters estimated. To decide which parameters may be constrained or dropped, we examine the standard errors of the parameters. Since underspecification is generally a larger danger than overspecification (cf. Mosteller & Tukey, 1977), parameters are dropped only when their two-sided p-value is larger than 0.10.

The maximal model is:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + u_{1j} X_{ij} + u_{0j} + e_{ij} \qquad (3.26)$$

[1]VARCL does not compute a standard error for σ^2. For the higher level variances, it computes the standard error of their square root, designated as 'sigma' by the program.

With four pupil level and three class level explanatory variables, this model includes one pupil level variance σ^2, the five class level variances of $ß_0$ to $ß_4$, and the 10 covariances between $ß_0$ and $ß_4$. Thus, the class level covariance matrix Σ contains 15 parameters to be estimated.

At the start of the analysis, when the maximal model is defined, VARCL asks whether the covariances between the random regression slopes in the matrix W should be fixed at zero. If this option is chosen, VARCL estimates a much simpler model. In many analyses, the covariances between the slopes turn out to be very small, and fixing them to zero speeds up the computations. There are also statistical advantages: we are estimating fewer random parameters, and the resulting models are generally more stable. HLM and MLn include all covariances by default in the model. For comparison, I start the VARCL analysis with a maximal model in which all explanatory variables are included and in which all covariances are assumed to be non-zero (model 1). In the next model (model 2), I fix all slope-by-slope covariances to zero, estimating only slope-by-intercept covariances. This results in a class level covariance matrix Σ that contains only nine parameters: the five variances and the four covariances of the four slopes with the intercept. I use the overall test on the deviances to test whether fixing all slope by slope covariances at zero is justified. The deviance of the maximal model with all slope by slope covariances estimated is 868.1. The deviance of the maximal model with all slope by slope covariances fixed at zero is 869.3, and the difference with the deviance of the previous model is 1.2. A comparison of the parameter count shows that we have restricted six covariances to zero. Thus, the value of 1.2 is a chi-square variate with six degrees of freedom, which gives a p-value of 0.98. I conclude that these covariances may be fixed at zero.

Next, the model is simplified by fixing random regression slopes or by dropping variables that are clearly not significant. The results (deviances and associated tests) of this model exploration by backward elimination of nonsignificant effects are summarized in Table 3.7.

If we compare model (1) and (2), it turns out that the slope-by-slope covariances can be fixed at zero, leaving only intercept-by-slopes covariances in the model. The difference in deviance between models (5) and (6) corresponds to a chi-square of 6.0 with two degrees of freedom, which is of borderline significance

(p=.05). This indicates that dropping the regression slopes for `repeat' and `teacher gender' results in a significant reduction of the model fit. The other p-values are clearly non-significant. Since the difference between model (5) and model (6) is of borderline significance but rather small, I could choose to use model (6) because it is the most parsimonious model. Instead, I prefer to keep model (5) as the most parsimonious model that describes the data well.

Table 3.7. Summary of VARCL models for example data

				difference with previous model		
	Model:	Deviance	#Parameters	Chi²	df	p
(1)	Max, Slope×Slope	868.1	24	-	-	-
(2)	Max, I×S only	869.3	18	1.2	6	.98
(3)	Fix lonely pretest	870.8	16	1.5	2	.47
(4)	Fix repeat & ethn.	875.8	12	5.0	4	.29
(5)	Drop teacher courses	876.5	11	0.7	1	.41
(6)	Drop rept & t. gend.	882.5	9	6.0	2	.05

Model (5) corresponds to the single model equation:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \\ + u_{0j} + u_{1j}X_{1ij} + e_{ij}. \qquad (3.27)$$

The parameter and variance estimates for this model are presented in Table 3.8 below:

Table 3.8 Estimates for the example data, model (5)

Variable:		γ	SE	p	
intercept		.25			
pup. gender	-.17	.08	.05		
repeat		.15	.08	.09	
ethnic		.13	.07	.05	
lonely pretest	.63	.04	.00		
teach. exper.	-.03	.01	.04		
		Variance	Sigma	$SE_{(sigma)}$	
pupil level (σ^2):	.39				
intercept (σ_{00}):	.14	.37	.06	(p=0.00)	
slope p.gend.(σ_{11})		.08	.28	.09	(p=0.00)

In model (5), the intercept variance (σ_{00}) is much lower than in the `intercept only' model. As before in the HLM analysis, this is difficult to interpret, because the model also contains a random slope for a pupil variable, and the two estimates are correlated. Again, simple linear transformations of this explanatory variable may change this variance.

The standard deviation for the slope of pupil gender (Sigma in Table 3.8) is 0.28. This is large, compared to the mean value of -0.17. This means that it is likely that at least in a few classes the slope of pupil gender is positive instead of negative (we can confirm this by requesting their estimated values, which VARCL calls `posterior means', and inspecting these.) The implication is that, while in general girls are less lonely than boys, there are a few classes where they actually are more lonely than boys.

Since there are three teacher variables, we can investigate three different interactions with pupil gender to explain the slope variation. The model which includes all three interaction variables has a deviance of 872.7, which differs from model (5) by 3.8. Since we have added three interaction variables to the model, this difference in deviance is referred to the chi-square distribution with three degrees of freedom. The resulting p-value is 0.28, which suggests that adding the interactions has not improved the fit of the model. However, only the slope for the interaction between pupil gender and teacher experience is larger than its standard error. The next model includes only this interaction; this model has a

deviance of 873.1. The difference between the deviance of this model and the deviance of the model that includes all three interactions is 873.1-872.7=0.4, with two degrees of freedom; this shows that a model with one interaction fits as well as a model with all three interactions. The difference between the deviance of the model with one interaction and the deviance of model (5), which has no interactions, is 876.5-873.1=3.4, at the cost of including one more parameter in the final model (the regression coefficient of the interaction variable); a chi-square test with one degree of freedom yields a p-value of 0.06. This is of borderline significance, and I decide to keep this interaction in the model. The parameter and variance estimates for this final VARCL model are in Table 3.9:

Table 3.9 Final estimates for the example data

Variable:	γ	SE	p	
intercept	.25			
pup. gend.	.15	.18	.41	
repeat	.15	.08	.08	
ethnic	.13	.07	.06	
lonely pretest	.64	.04	.00	
teach. exper.	-.02	.01	.28	
p.gend.×t.exp.	-.03	.01	.06	
	Variance	Sigma	$SE_{(sigma)}$	
pupil level (σ^2):	.39			
intercept (σ_{00}):	.14	.38	.06	(p=0.00)
slope p.gend.(σ_{11})	.05	.24	.09	(p=0.01)

If we compare these results with the previous model in Table 3.8, we note that after the introduction of the interaction the sign for the overall effect of pupil gender has changed, and that the p-values for pupil gender and teacher experience are no longer significant. Furthermore, adding the interaction variable has diminished the variance of the slope for pupil gender, but there is still significant unexplained variation left. Apart from the pretest, no variables are significant at the conventional alpha level of 0.05. Since the standard errors and significance values are approximate, and some of these coefficients are of borderline significance, the conclusion is that these variables merit further investigation with a larger sample.

To interpret the interaction, we must compute the regression slopes of teacher experience for boys and girls separately, just as we did in the previous analysis with HLM. Pupil gender is scored 1 for boys, 2 for girls, and teacher experience is measured simply in years. Ignoring other variables, the regression equation relating posttest loneliness to teacher experience can be found by taking the appropriate part of the regression equation (from Table 3.9) and filling in the values for pupil gender and teacher experience. The equations linking posttest loneliness to teacher experience (Exp.) are for the boys: (Y=1*0.15-0.02*Exp.-1*0.03*Exp.=) Y=0.15-0.05*Exp.; and for girls (Y= 2*0.15-0.02*teach.exp.-2*0.03*Exp.=): Y=0.30-0.08*Exp. In our analysis, for teachers with less than five years of experience, the girls are more lonely than the boys, and for teachers with six or more years of experience, the girls are less lonely than the boys. Since in our data set the average teacher experience is more than 12 years, much larger than the turnover point of five years, the regression slope for `pupil gender', *not* taking teacher experience into account, is -0.17 (see Table 3.8), just the opposite of what one might expect from the positive value of 0.15 for the coefficient of pupil gender in the last model. With the HLM analysis we found basically the same results, reminding us that it is dangerous to interpret main effects in the presence of significant interactions, and that we should interpret interactions with reference to the range of values of the explanatory variables actually present in the data.

3.3 MLn Analysis of the Example Data

MLn differs from HLM and VARCL in that it can analyse data with an arbitrary number of levels (assuming sufficient computer memory), and offers analysts a choice between Full maximum Likelihood estimation (called IGLS) and Restricted Maximum Likelihood estimation (called RIGLS). Just as its precursor ML3, it has many built-in commands for data-manipulations such as centering or standardizing variables, and graphical commands to produce various plots.

I repeat the basic hierarchical regression model with one variable X_{ij} at the pupil

level, and one variable Z_j at the class level:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} Z_j X_{ij} + u_{1j} X_{ij} + u_{0j} + e_{ij}$$

The MLn notation is essentially equal to the notation I have used so far. There is one interesting difference. For each group, MLn estimates a micro level regression equation of the form:

$$y_{ij} = ß_{j0} X_{ij0} + ß_{j1} X_{ij1} + \ldots + ß_{jP-1} X_{ijP-1} + e_{ij} \qquad (3.28)$$

where y_{ij} is the outcome variable for individual *i* in group *j*
X_{ijp} are the individual level variables
e_{ij} is random error, and
$ß_{jp}$ are the random regression coefficients within group *j*

As equation (3.28) makes clear, MLn treats the intercept coefficient exactly the same way as the slope coefficients. Generally the explanatory variable X_{j0} will have the value `1' for all cases, which makes $b_{j0}$ the usual intercept. VARCL does precisely this; the program automatically assigns the value `1' to X_{j0}, and for that reason we simplified the VARCL section by omitting X_{j0}. MLn does not automatically assign any value to the explanatory variable X_{j0}; users must do that themselves. As a rule, users will set the value of the explanatory variable X_{j0} to `1', but it is perfectly possible to use other values than `1' or to use a variable instead of a constant. This is one of the features that make it possible to use MLn to analyze all sorts of nonstandard models.

There is a second intriguing difference between MLn and the other two programs. With HLM and VARCL, the regression slopes at the highest level (group level) are always fixed. MLn allows all regression coefficients to be random at all levels. The variances and covariances of the beta's are contained in the covariance matrix Ω (omega). Since the regression coefficients may be random at all available levels, there are covariance matrices: Ω_1, Ω_2, ... Ω_l, corresponding to the number of levels specified for MLn. The higher level covariance matrices Ω_2, Ω_3, etcetera, have the same interpretation as the matrix Tau in HLM and Sigma

in VARCL; they hold the higher level (co)variance components of the regression coefficients. Ω_1 is different; it is defined on the first (lowest) level, and can be used to model heterogeneous variances on the first level (Prosser et al., 1990, explain how to do this in the MLn manual). This is a feature only MLn has, and again this makes it possible to use MLn to analyze many nonstandard models. For instance, Goldstein (1994) uses this feature to model a cross-classified structure, and Van Duijn, Snijders and Lazega (Van Duijn, Snijders & Lazega, 1994) use it to model sociometric network data.

The higher level variance of the regression coefficients (intercept and slopes) may again be explained by a between group model. In the MLn notation, this equation takes the form:

$$ß_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j}. \qquad (3.29)$$

Apart from the covariance matrices Ω, for which I use Σ, the MLn notation is identical to the notation in this book.[1]

MLn requires only one data file, which has to be sorted, and must contain variables that identify the units at all levels used. MLn is a totally interactive program. The data are kept in a worksheet, where the variables define the columns and the cases define the rows.[2] There are many commands that allow the user to `play around' with the data, such as standardizing variables or centering variables around either the grand mean or the group means. MLn also has powerful graphic commands to produce overall or group-wise plots. Together with the availability of macro routines, which let users automatically repeat computations or use residuals from one analysis automatically as input in another

[1]When there are more than two levels, the Σ should carry a subscript to indicate to which level it belongs. As long as there is no risk of confusion, I will use the simpler notation.

[2]All data must reside in RAM memory. As a consequence, the computer must have enough memory to hold all the data, including storage for a number of variables used internally by the program. The example data fit in a PC/AT version of ML3, including predicted values and residuals for plots. However, it did strain the machine, and a larger data set would require a 486 PC and the extended memory version of the program. MLn exists needs extended memory.

model, this makes MLn a very powerful program. It also makes the program fairly complicated, and only experienced `power users' will use all its features.

For the simple purpose of our example, the most interesting feature of MLn is the fact that it is interactive and gives users complete control over the computations. This makes it easy to try out different subsets of explanatory variables and different error structures. It is also not necessary to wait until all the computations are completed. MLn allows analysts to perform the iterative computations step by step, or only a few iterations at a time. If inspection of the preliminary results indicates that the model is obviously wrong, it can be changed at any moment.

MLn offers a choice between Full Maximum Likelihood estimation (called Iterative Generalized Least Squares, or IGLS; comparable to the estimation method used in VARCL) and Restricted Maximum Likelihood estimation (called Restricted Iterative Generalized Least Squares or RIGLS; comparable to the estimation method used in HLM). Since IGLS is generally faster and numerically more stable, it is the preferred method to start with.

Fitting a model to our example data without explanatory variables gives us the by now familiar decomposition of the variance between level 1 (the individual level: 0.72) and level 2 (the class level: 0.31). Just like VARCL, MLn produces parameter estimates and their standard errors, but no p-values. MLn also includes commands to calculate the probability levels of various test statistics, so it is easy to calculate whether a certain effect is statistically significant (bearing in mind that, as with the other programs, the standard errors are valid only for large samples).

The first model I examine includes all pupil variables in the model, with those teacher variables that have an (approximate) p-value of less than 0.10, and only the intercept coefficient random.

The pupil level model is:

$$Y_{ij} = ß_{0j} + ß_{1j}X_{1ij} + ß_{2j}X_{2ij} + ß_{3j}X_{3ij} + ß_{4j}X_{4ij} + e_{ij} \qquad (3.30)$$

and the class level model is:

$$ß_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} \quad (3.31a)$$
$$ß_{1j} = \gamma_{10} \quad (3.31b)$$
$$ß_{2j} = \gamma_{20} \quad (3.31c)$$
$$ß_{3j} = \gamma_{30} \quad (3.31d)$$
$$ß_{4j} = \gamma_{40} \quad (3.31e)$$

which gives

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \\ + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} + e_{ij}. \quad (3.32)$$

In this model only the intercept is allowed to be random; all other regression coefficients are fixed. The parameter estimates are in Table 3.10:

Table 3.10 MLn model with all pupil and teacher variables

	gamma	SE	p	
intercept	.23		.19	.21
pup. gender	-.15		.06	.02
repeat	.12		.09	.15
ethnic	.14		.07	.03
lonely pretest	.64		.04	.00
teach. gndr .31		.16	.05	
teach. exper.	-.03		.01	.03
	Variance			
pupil level (σ^2)	.41		.03	.00
class level (σ_{00})	.14		.04	.00

At this point, I decided to investigate the random structure for the regression slopes further, and to use the interactive nature of MLn to speed up the process. One by one, each of the four pupil variables was made random at level 2; in other words, its slope was allowed to vary between groups. In terms of equation (3.31), this means that the corresponding ß is assumed to be random, and the appropriate error term u is added to equation (3.31b) to (3.31e). After only two iterations, the variance estimate σ_{pp} was inspected, with its standard error and the estimate at the previous iteration. Although convergence was not reached in any of the

models, the results were absolutely clear. Only the between group variance σ_{11} of the slope of the pupil variable `gender' approached significance. All other variances were estimated as approximately zero with large standard errors, and they were therefore dropped from the model. The between groups covariance $\sigma_{10}$ between the regression slope of `pupil gender' and the intercept was also small (-0.03, with a standard error of 0.04), and this covariance term was also dropped (note that this is different from the previous HLM and VARCL analyses, where this covariance term is always kept in the model by program default, and there is no way to drop it). This results in the following model:

At the pupil level:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \beta_{4j}X_{4ij} + e_{ij}. \qquad (3.32)$$

and at the class level:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + u_{0j} \qquad (3.33a)$$
$$\beta_{1j} = \gamma_{10} + u_{1j} \qquad (3.33b)$$
$$\beta_{2j} = \gamma_{20} \qquad (3.33c)$$
$$\beta_{3j} = \gamma_{30} \qquad (3.33d)$$
$$\beta_{4j} = \gamma_{40} \qquad (3.33e)$$

which gives

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \\ + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{12}X_{1ij}Z2j + u_{0j} + u_{1j}X_{1ij} + e_{ij} \qquad (3.34)$$

The variance/covariance matrix Σ at the class level is:

$$\Sigma = \begin{matrix} \sigma_{00} & - \\ - & \sigma_{11} \end{matrix}$$

There is no covariance term σ_{01}, since I fixed that at zero. The results from this model are in Table 3.11:

Table 3.11 Simplified MLn model with pupil and teacher variables

	gamma	SE	p	
intercept	.20		.19	.28
pup. gender	-.16		.06	.05
repeat	.14		.09	.10
ethnic	.13		.07	.06
lonely pretest	.63		.04	.00
teach. gndr .29		.16	.07	
teach. exper.	-.02		.01	.08
	Variance			
pupil level (σ^2)	.39		.03	.00
class level (σ_{00})	.13		.04	.00
class level (σ_{11})	.07		.04	.00

Since the between class variance of the regression slope for pupil gender approaches the conventional significance level (p=0.06), we may try to model this regression slope with the class level variable `teacher experience', which we also used in the HLM and VARCL analyses. This expands equation (3.33b) to:

$$ß_{1j} = \gamma_{10} + \gamma_{12}Z_{2j} + d_{1j} \qquad (3.35)$$

and equation (3.38) now becomes:

$$Y_{ij} = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{30}X_{3ij} + \gamma_{40}X_{4ij} + \gamma_{01}Z_{1j} + \gamma_{02}Z_{2j} + \gamma_{12}X_{1ij}Z_{2j} + u_{0j} + u_{1j}X_{1ij} + e_{ij} \qquad (3.36)$$

To estimate the model given by equation (3.36), we must compute the interaction variable $X_{1ij}Z_{2j}$. This is similar to the procedure with VARCL, but this time the data manipulation commands of MLn can be used to compute the needed interaction variable, without leaving the program. The regression slope for the interaction is -0.03 with a standard error of 0.01, which is significant at the 5 percent level (p=0.05). Since I use the (default) IGLS estimation procedure, which includes both the random part and the fixed part of the model in the likelihood function, I can test the significance of the interaction parameter γ_{12} by comparing the values of the likelihood function for the models with and without the

interaction variable. MLn can compute the deviance, and the difference between the two deviances can be treated as a chi-square. The deviance for the model with the interaction variable is 873.4, while the deviance of the model without the interaction equals 877.1. The difference of 3.7 can be referred to the chi-square distribution with one degree of freedom, which gives a p-value of 0.05. The results of the final MLn model are in Table 3.12. Since MLn allows us to switch at will between IGLS and RIGLS estimation, the final model was also estimated with RIGLS. In this particular case, both methods produce almost identical results.

Table 3.12. MLn, pupil and teacher variables, pupil gender random

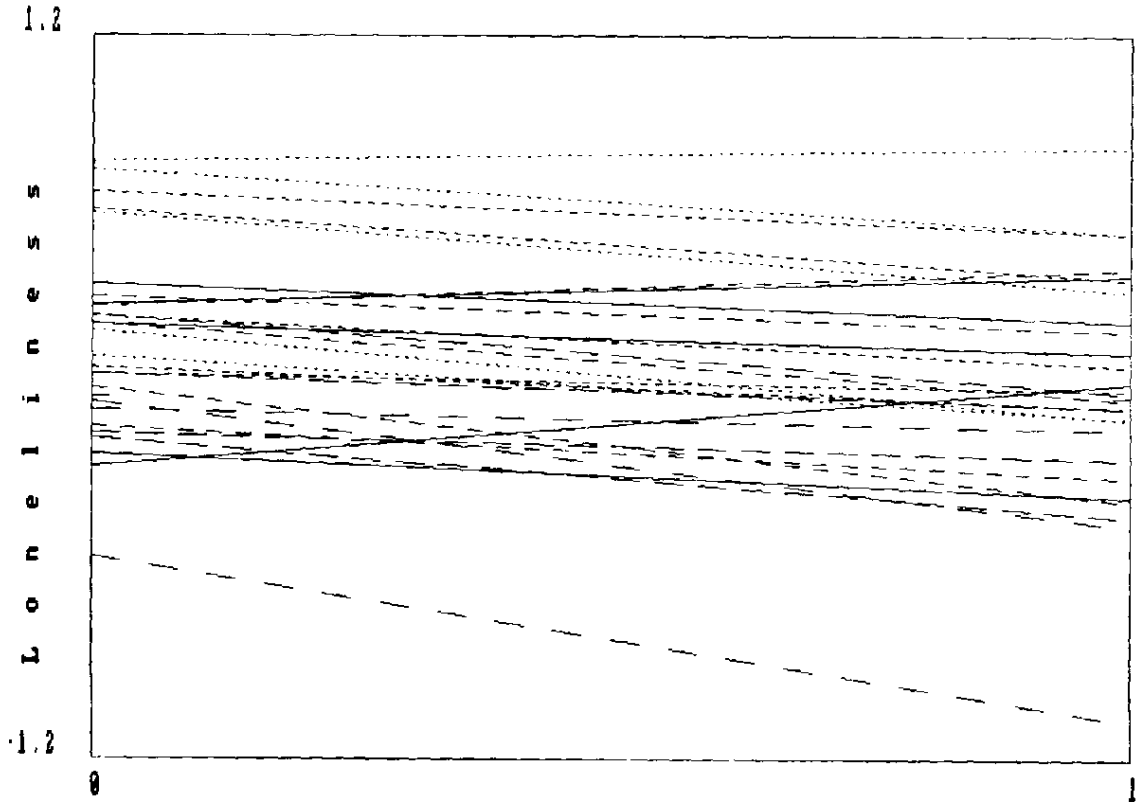
	IGLS:			RIGLS:		
	gamma	SE	p	gamma	SE	p
intercept	.10	.20	.60	.11	.21	.60
pup. gender	.16	.18	.36	.16	.18	.37
repeat	.14	.09	.10	.14	.09	.09
ethnic	.12	.07	.07	.12	.07	.07
lonely pretest	.64	.04	.00	.64	.04	.00
teach. gndr .29	.16	.07	.29	.17	.09	
teach. exper.	-.02	.01	.27	-.02	.02	.30
p.gendr.×t.exp.	-.03	.01	.05	-.03	.01	.05
	variance			variance		
pupil level (σ^2)	.39	.03	.00	.39	.03	.00
intercept (σ_{00})	.13	.05	.00	.15	.05	.00
pup.gndr(σ_{11})	.05	.04	.11	.06	.04	.07

Adding an interaction variable has diminished the variance of the slopes for pupil gender. As in the previous analyses with HLM and VARCL, the multicollinearity between pupil gender, teacher experience, and the interaction variable results in changes in the signs of the regression coefficient for pupil gender and a large standard error which makes that coefficient not significant. In MLn, it is easy to check for multicollinearity by using the built-in statistical procedures to calculate the correlations between the explanatory variables. The correlation between `pupil gender' and `teacher experience' turns out to be low (r=0.13). The correlation between `pupil gender' and the interaction variable is 0.50, which is considerable,

and the correlation between `teacher experience' and the interaction variable turns out to be 0.84, which is high.

As we just noted, if we add the interaction to the model, the direct effect of pupil gender is no longer significant. To interpret both the interaction and the direct effect, it is again helpful to compute the regression equation for teacher experience for boys and for girls separately. After the similar calculations in the HLM and VARCL analyses, I leave this as an exercise for the reader.

A strong point of MLn are its powerfull graphics commands. The analyses suggest that the effect of pupil gender varies across the classes. With MLn, we can plot the 28 slopes for `pupil gender' in one plot and inspect them visually. The result is on the next page. The regression coefficient for `pupil gender' in Table 3.11 is -.16, which means that in general girls are more lonely than boys. The variance of this regression coefficient across the classes is 0.39, which corresponds to a standard deviation of 0.62. Thus, for most classes the slopes should go down, but a few might even go upwards. If we inspect Figure 3.1 on the next page, the generel view is *not* that all slopes are basically going down from left to right. Instead, the most striking impression is that we have one class that stands out from the rest, with a lower intercept and a slope that clearly goes down. The other slopes do not show a prominent trend. Judging from Figure 3.1, it is quite possible that the significant interaction effect is actually caused by one single outlier. MLn has the capability to identify this outlier: it is class number twelve. Before we draw strong conclusions about the different effect of pupil gender in classes with teachers of varying experience, it would be wise to inspect the data more carefully. It is likely that without class twelve, the data would show no random slope variation. There may be something unique about class number twelve, which is not captured by the available explanatory variables. Whatever our conclusion, Figure 3.1 illustrates the value of a graphic inspection of our data. MLn contains many more graphical commands, for examples see e.g., Goldstein (1987, 1995) and Prosser et al., (1991) and Woodhouse (1995).



Pupil gender (0=boy, 1=girl)

Fig. 3.1 Slopes for pupil gender in all classes

4. Special Applications of Multilevel Regression Models

4.1 Multilevel Models for Meta-analysis

Meta-analysis, or integrative analysis, as it is often called, is a quantitative approach to reviewing the research literature. The primary goal of meta-analysis is to generalize from a set of studies about a specific substantive issue, by statistically combining quantitative study outcomes from existing research on a particular question (Jackson, 1980). The basic idea is to apply formal statistical methods to the results of a specific set of studies. This statistical approach is one of the main characteristics that distinguish meta-analysis from the more traditional narrative literature review (Bangert-Drowns, 1986). If the results of the studies do not differ too much, we apply statistical procedures to combine all the results into one average outcome. If the results of the studies vary a lot, the primary goal of meta-analysis becomes to answer the question why the results vary. Thus, when results vary, the analyst attempts to explain the different results as the consequence of differing study characteristics (such as type of design or subjects used).

In a recent edition of their classical book on meta-analysis, Hunter and Schmidt (1990) present an instructive example to show why a formal statistical analysis of study outcomes has its advantages. In this example Hunter and Schmidt (1990, p24) present the results of 30 hypothetical studies of the correlation between `Organizational Commitment' and `Job Satisfaction'. The sample size in the 30 studies varies, and so does the correlation reported as the main result of the studies. Nineteen of the 30 studies report a significant positive correlation, nine report a non-significant positive correlation, and two report a non-significant negative correlation. There are six known background characteristics of the sample of persons used in a particular study: sex, organization size, white vs. blue collar jobs, race (white, black, or mixed), average age, and geographical location (north vs. south). In a narrative review of the results, Hunter and Schmidt ask the question `Why are commitment and

satisfaction correlated within some organizations and not within others?', and proceed to interpret the different study outcomes by relating them to the known background characteristics of the people within the different organizations. Their interpretations sound fairly convincing.

As Hunter and Schmidt carefully explain, all these `interpretations' are completely spurious, since the 30 study outcomes were actually all generated from one single population characterized by a population correlation coefficient of r=0.33. For each study the sample size was chosen randomly, and the sampling error of the correlation in that study was taken from a random distribution with a sampling variance appropriate to the study's sample size. In other words, all the interesting looking variation among the 30 studies in this example is entirely due to sampling variation. Presumably, if a statistically correct method were used to meta-analyze the 30 study outcomes, an analyst should come to the conclusion that, within the limits of sampling variance, all studies actually report the same result.

Clearly, in a meta-analysis the most important preliminary question is, whether the results differ more from each other than corresponds to the random sampling variation that is expected given the studies' sample sizes. If the results do not differ more than is expected given the pure sampling error, they are called *homogeneous*, meaning that they come from a single population. In the next analysis step we would want to estimate the common value of the population parameter of interest. If the results differ more than expected given the pure sampling variation, they are called *heterogeneous*, meaning that they come from different populations. In this case, estimating the `average' result is not the primary goal; instead, our goal becomes to analyze the excess variation as a function of the known study characteristics such as the age or sex composition of the sample, or methodological characteristics such as the methodological quality of the study.

There are various methods to analyze and combine separate study results. Hunter and Schmidt (1990) present many methods to correct study results for sampling variance and other potential sources of bias. Hedges and Olkin (1985) provide the most statistically thorough discussion to date of the problems associated with statistically integrating research findings, and discuss procedures

for many statistics (means, proportions, correlations.)

Table 4.1 below presents the results of the 30 studies, with the Fisher Z transformation and the standard errors.

Table 4.1 Hypothetical results for 30 studies on job satisfaction

Study#	r	Z	se(z)	sex	size	collar	race	age	south
1	.46	.49	.06	0	0	1	0	0	2
2	.32	.33	.01	1	1	0	1	1	2
3	.10	.10	.04	1	1	1	2	2	2
4	.45	.48	.04	1	1	1	2	1	2
5	.18	.18	.01	0	1	0	2	2	2
6	.40	.48	.02	0	0	0	2	0	2
7	.56	.63	.05	1	0	0	1	0	0
8	.41	.43	.02	0	1	1	2	1	0
9	.55	.61	.05	0	0	1	0	0	2
10	.44	.47	.02	0	0	0	2	0	2
11	.34	.35	.02	1	1	0	2	1	2
12	.33	.34	.02	1	0	0	2	0	2
13	.14	.14	.05	1	0	1	0	2	0
14	.36	.37	.06	1	0	1	2	2	2
15	.54	.60	.04	0	1	1	2	1	0
16	.22	.22	.04	1	0	0	2	1	0
17	.31	.32	.02	0	1	0	2	1	2
18	.43	.45	.02	0	1	0	2	1	2
19	.52	.57	.06	1	0	0	2	1	0
20	-.10	-.13	.02	1	0	1	2	2	2
21	.44	.47	.02	0	1	0	1	1	2
22	.50	.54	.05	0	0	1	2	1	0
23	-.00	-.02	.06	1	0	1	0	2	0
24	.32	.33	.02	1	1	1	2	1	1
25	.19	.19	.06	0	0	1	0	2	2
26	.53	.59	.04	0	0	0	0	0	0
27	.30	.30	.02	1	1	1	2	1	0
28	.26	.26	.05	1	0	1	2	0	0
29	.09	.09	.04	0	0	0	2	2	2
30	.31	.32	.04	0	0	1	1	0	0

The problem of combining the varying results from different studies has some similarity to the multilevel problem of combining the varying micro-models from different groups or contexts. In the example of the 30 studies of Hunter and Schmidt, the contexts are the 30 studies, each of which has its unique individual characteristics. If we had access to the original data of all the studies, we could analyze them using the hierarchical regression model. But in meta-analysis we generally do not have access to the raw data. Still, the statistical problem looks familiar. In multilevel modeling we have a number of regression models computed in different contexts, and we want to estimate the expectation and the variability of the various regression coefficients, and draw conclusions based on all available information. In meta-analysis we have a number of statistics (in our example: correlations) computed in different contexts, and we want to assess their average value and their variability, and again draw conclusions based on all available information.

The similarity is more than superficial. Raudenbush and Bryk (Raudenbush & Bryk, 1985, 1987; Bryk and Raudenbush, 1988, 1992) have pointed out that meta-analysis may be viewed as a special case of the two level hierarchical linear model. In each study, a within study model is estimated, and a second level or between study model is added to explain the variation in the within study parameters as a function of differences between the studies. This is completely analogous to the example discussed earlier, where a within class (pupil level) model is estimated in each class, and a between class model is added to explain the variance of the class level regression coefficients. The variability within the studies is considered to be sampling variability, which is known if the relevant sampling distribution and sample size are known (since the variance is assumed to be known Raudenbush & Bryk (1993) call this the V-known model.) The variability between the studies reflects both sampling variance and systematic differences between the results of different studies. If the study level variance is significant, the studies' results are assumed to be *heterogeneous*, meaning that there are indeed systematic differences between the studies. If the study level variance is not significant, they are assumed to be *homogeneous*, meaning that the apparent differences between the studies are just sampling variance. Raudenbush and Bryk (1992) show that the computational procedure incorporated in HLM are useful as

a general approach to meta-analysis.

To use HLM for meta-analysis, we need for each study the value of the result we want to analyze, and its sampling variance in that particular study. Suppose that we want to meta-analyze the 30 correlations from Hunter and Schmidt's example. Since the sampling distribution of correlations is not normal, we first transform the observed correlations r to standard normal Fisher Z variates z using the Fisher-Z transformation: Z=0.5*ln((1+r)/(1-r)) (see Hays, 1973). The sampling variance of a Fisher Z is known, it is $v_z = 1/(n-3)$ where n is the sample size on which the corresponding correlation coefficient is based.[1]

Hence, our within study model is:

$$z_j = \zeta_j + e_j \qquad (4.1)$$

where ζ_j is the true parameter value of the Fisher Z in study j, and e_j is the sampling error in study j. The e_j are assumed to be normally distributed with variance σ_j^2; thus z_j varies randomly between studies as a result of the joint effect of sampling error and real parameter variance. Note that the sampling variances e_j are independent but not identically distributed; they are assumed to have a normal distribution with a variance σ_j^2 that is different for each study because it depends on the study's sample size.

The between studies model for the random parameter ζ_j is:

$$\zeta_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + .. + u_j \qquad (4.2)$$

where the Z_j's are study level variables, such as the sample composition or the geographical location. The γ's are study level fixed regression coefficients, and u_j is the study level random error assumed to be distributed normally with variance σ^2_u. If we substitute equation (4.2) into (4.1), we obtain:

---

[1] A well-known effect size measure is the standardized mean difference $d=(M_E-M_C)/S$, with sampling variance $(N_E+N_C)/(N_E N_C)+d^2/(2(N_E+N_C))$. For a meta-analysis, one would substitute the estimated value for d in the equation for the sampling variance, and proceed in the same manner as descibed in 4.1 for the Fisher Z's. See Bryk and Raudenbush (1992) and Hedges and Olkin (1985) for other effect size measures and their sampling variances.

$$z_j = \gamma_0 + \gamma_1 Z_{1j} + \gamma_2 Z_{2j} + .. + u_j + e_j \qquad (4.3)$$

In equation (4.3) the variance σ_j^2 of e_j is the sampling variance, and the variance σ^2_u of u_j is the parameter variance. Just as in a conventional multilevel analysis, HLM gives us estimates of the fixed parameters γ_i and of the sampling variance and the between studies parameter variance, together with the associated significance levels. Since we know that the 30 correlation coefficients in our example are actually homogeneous, HLM ought to tell us that the parameter variance between the studies is zero. For the preliminary analysis, we leave out the study characteristics, and estimate the model which is analogous to the `intercept only' model, which is given by:

$$z_j = \gamma_0 + u_j + e_j \qquad (4.4)$$

If the parameter variance is very small, the convergence of HLM may be slow. This turned out to be true; the algorithm needed 558 iterations to converge.[1] HLM estimates γ_0 as 0.35, with a standard error of 0.03, and σ_u is 0.0007, with an associated p-value (by a chi-square test) of p=0.40. The large (non-significant) p-value for the parameter variance σ_u indicates that there is no significant parameter variance; all observed variation is sampling variation. HLM reports the same information in the form of a reliability estimate for the regression coefficient γ_0, which is estimated as 0.02.

Since there is no reliable parameter variance between the studies, we may conclude that the results are homogeneous, and it makes no sense to use the six study level variables to attempt to explain this nonexisting variation. If there were reliable parameter variance, we would conclude that the results are heterogeneous, and in the next analysis step the study level variables can be included in the model to explain this parameter variance in a manner completely analogous to conventional multilevel analysis. This analysis would examine directly the possible causes of the differences in results, another important goal of

[1]Since each iteration was very fast, the total computing time of the V-known (meta-analysis) version of HLM was still reasonable: a few minutes on a PC/AT).

meta-analysis. The article by Raudenbush and Bryk (1987) is a clear example of the kind of reasoning involved.

In our example, there is one more step. Since all variation is sampling variation, the study results may be regarded as homogeneous, and the value of 0.35 is a useful estimate of a common Fisher Z value z for all studies. Since we are actually interested in correlations, we have to translate this result back into the corresponding correlation by the inverse of the Fisher transformation, which is given by: r=(exp(2*Z)-1)/(exp(2*Z)+1). This produces a common value for the correlation of r=0.34, which is very close to the known population value of r=0.33.

In general, we should keep in mind that the decision that a specific set of results is homogeneous or heterogeneous, is a statistical decision given a certain significance level, and consequently we may commit an error. When the number of studies is small, the power of the test for heterogeneity may not be good. It is possible to find one specific significant study level variable, while there is no significant overall heterogeneity. The procedure outlined above would apply to an explorative search for relevant study level variables. If we have a strong theory that predicts that certain study level variables are important, we should test them even if the preliminary analysis shows that there is no significant between-study variation.

Meta analysis can also be performed using MLn. Since the variance at the lowest level is assumed known, the MLn model must exclude the usual constant from the random part of the lowest level, and in its place include the standard error as a predictor in the random part of the lowest level, with its associated coefficient constrained to be equal to one (cf. Lambert & Abrams, 1995). A problem with this approach is that using MLn the significance test for the parameter variance will usually employ on the Wald test, dividing the variance estimate by its standard error and refering the result to the standard normal distribution. Bryk and Raudenbush (1992) argue against this test because the sampling distribution of a variance is not normal, and both HLM and VKHLM employ a different method. If MLn is used for meta-analysis, the difference in test procedures is a potential source of confusion. The reason is that the conventional homogeneity tests used in meta-analysis (Cf. Hedges & Olkin, 1985) are a special case of the chi-square test used in HLM. If VKHLM is used to estimate the

intercept only model, the result of the chi-square test for the parameter variance will be virtually equal to the equivalent chi-square test in the random-efects model described by Hedges and Olkin (1985).[1] The result of the Wald test used in MLn can be widely different, especially when the number of studies is not large. If MLn is used for meta-analysis, I recommend to use its built-in calculation commands to implement the chi-square test described by Bryk and Raudenbush (1992, p. 163).

4.2 Non-normal Data; the Analysis of Proportions

The models discussed so far assume a continuous dependent variable and a normal error distribution. If the dependent variable is a scale in which the responses to a large number of questions are summated to one score, the data generally approximate normality. However, there are situations in which the assumption of normality is clearly violated. For instance, in cases where the dependent variable is a single dichotomous variable, both the assumption of continuous scores and the normality assumption are obviously untrue. If the dependent variable is a proportion, the problems are less severe, but both the assumption of continuous scores and normality are still violated. Also, in both cases, the assumption of homoscedastic error is violated.

The classical approach to the problem of non-normally distributed variables and heteroscedastic errors is to apply a transformation to achieve normality and reduce heteroscedasticity, followed by a traditional analysis with ANOVA or multiple regression. Some general guidelines for choosing a suitable transformation have been suggested for situations in which a specific transformation is often successful (cf. Kirk, 1968; Mosteller and Tukey, 1977). For proportions an appropriate transformation is the arcsine transformation: $f(x) = 2 \arcsin \sqrt{x}$, or the logit transformation: $f(x) = \ln(x/(1-x))$. When the dependent variable is a frequency count of events with a small probability, such as the number of errors made in a school essay, the data tend to follow a Poisson

[1]The models are identical, but the test results may diverge somewhat because the estimation methods are different.

distribution, which can often be normalized by taking the square root of the scores: $f(x) = \sqrt{x}$. When the data are highly skewed, which is usually the case if, for instance, reaction time is the dependent variable, a logarithmic transformation is often used: $f(x) = \ln(x)$, or the reciprocal transformation: $f(x)=1/x$. For reaction times the reciprocal transformation has the nice property that it transforms a variable with an obvious interpretation: reaction time, into another variable with an equally obvious interpretation: reaction speed.

The modern approach to the problem of non-normally distributed variables is to include the necessary transformation and the choice of the appropriate error distribution (not necessarily a normal distribution) explicitly in the statistical model. This class of statistical models is called *generalized linear models* (McCullagh & Nelder, 1983, 1989). Generalized linear models are defined by three components: 1) a linear regression equation, 2) a specific error distribution, and 3) a link function which is the transformation that links the predicted values for the dependent variable to the observed values. If the link function is the identity function ($f(x)=x$) and the error distribution is normal, the generalized linear model simplifies to ordinary multiple regression analysis. For other link functions and error distributions, the generalized linear model is estimated by complicated maximum likelihood procedures (cf. McCullagh & Nelder, 1983, 1989), but the results can be interpreted much as if they came from an ordinary linear model (cf. Aitkin et al., 1989, for examples).

Multilevel generalized models have been described by Wong and Mason (1985), Longford (1988, 1990), Mislevy and Bock (1989), and Goldstein (1991). For hierarchical regression models Longford has implemented several generalized linear models in VARCL. The basic algorithm is based on an iteratively reweighted least squares procedure, which is carried out together with the standard iterative procedure for the variance components (Longford, 1988, 1990). The link functions presently supported are the logistic link function for binary (dichotomous) and binomial data (proportions) (both are designated by the program as binomial data: dichotomous data are binomial data with only one trial), the logarithmic function for Poisson data, and the reciprocal link function for gamma distributed data. Comparable analyses can be carried out with MLn, using macro's to carry out the necessary transformations. Since the MLn

procedure is highly complicated, while the interpretation problems are analogous, I limit myself to an analysis of binomial data with VARCL.[1] The analysis presented below uses the logistic or *logit* link function; analyses using the logarithmic and the reciprocal link functions follow similar lines.

The example concerns data from a meta-analysis of studies that compared face-to-face, telephone, and mail surveys on various indicators of data quality (De Leeuw, 1992; for a more thorough analysis see Hox & De Leeuw, 1994). One of these indicators is the response rate; the number of completed interviews divided by the total number of eligible sample units. Overall, the response rates differ between the three data collection methods. In addition, the response rates also differ across studies, which makes it interesting to analyze what study characteristics account for these differences.

These meta-analysis data have a multilevel structure. The lowest level is the `condition-level,' and the higher level is the `study-level.' There are three variables at the condition level: the number of completed interviews in that specific condition, the number of eligible respondents in that condition, and a categorical variable indicating the data collection method used. The categorical data collection variable has three categories: `face-to-face', `telephone' and `mail.' To use it in the regression equation, it is recoded into two dummy variables: a `telephone-dummy' and a `mail-dummy.' In the `mail' condition, the mail-dummy equals one, and in the other two conditions it equals zero. In the `telephone' condition, the telephone-dummy equals one, and in the other two conditions it equals zero. The face to face condition is coded by a zero for both the telephone- and the mail-dummy. In this coding scheme the face-to-face condition is the baseline against which the two other conditions are compared.[2] There are three variables at

[1]Macro's allow users to implement their own solution. The basic procedure is described in Prosser et al., 1991. Goldstein (1994) descibes a new version of the ML3 macro that is more accurate than the previous version or the VARCL procedure used here. Since this new macro is rather complicated it is not used here. MLn has a more elaborate macro language. Also, the program MIXOR by Hedeker (1993) models ordinal data, which includes binary data as a special case.

[2]All coding schemes code a categorical variable with k categories into $k-1$ dummy variables. Cohen & Cohen (1983) and Kerlinger & Pedhazur (1973) discuss other coding schemes. Both VARCL and MLn have built-in provisions to generate dummy variables for categorical

the study level: the year of publication, the saliency of the questionnaire topic, and the methodological quality of the research. Most studies compared only two of the three data collection methods, a few compared all three. Omitting missing values, we have 45 studies in which a total of 99 data collection conditions are compared.

The dependent variable is the response rate. This variable is a proportion: the number of completed interviews divided by the number of eligible respondents. If we model these proportions directly by normal regression methods, we encounter two critical problems. The first problem is the fact that proportions do not have a normal distribution, but a binomial distribution, which (especially with extreme proportions and/or small samples) invalidates several assumptions of the normal regression method. The second problem is that a normal regression equation might easily predict values larger than 1 or smaller than 0 for the response rate, which are impossible values for proportions. Using the generalized linear (regression) model for the proportion p of potential respondents that are responding to a survey solves both problems, which makes it a more appropriate model for these data.

As I outlined above, the generalized linear model has three distinct components: 1) a linear regression equation, 2) a specific error distribution, and 3) a link function. The link function to be used for binomial data is the *logit* function, which is defined as logit(x) = ln(x/(1-x). The corresponding error function is the binomial distribution. Finally, using a logit regression model for the probability p allows us to use a linear regression model for the *logits* of the probabilities.

The hierarchical generalized linear model for our response rate data can be described as follows. In each group we have a number of individuals who may or may not respond. The population probability of responding is given by Π_{ij}, that is, for each individual r in each condition i of study j the probability of responding is the same. Note that we could have a model where each individual's probability of responding varies, with individual level covariates to model this variation. Then,

---

explanatory variables, with the first category (coded '1') as the baseline. Since dummy variables all derive from the same categorical variable, they are not independent. For this reason, when the effect of a categorical variable is assumed to vary between groups, all corresponding dummy variables must be declared random, and their slope-by-slope covariances should not be set to zero (cf. Longford, 1990).

we would model this as a three-level model, with binary outcomes at the lowest (individual) level. Since in this meta-analysis example we do not have individual data, the lowest level is the condition-level, with conditions (data collection methods) nested within studies.

Let P_{ij} be the observed proportion respondents in condition *i* of study *j*. While P_{ij} has a binomial distribution, *logit* (P_{ij}) has a distribution that is approximately normal (McCullagh & Nelder, 1989).[1] Thus, at the lowest level, we use a linear regression equation to predict logit (P_{ij}). The simplest model, corresponding to the intercept-only model in ordinary multilevel regression analysis is given by:

$$\text{logit}\,(P_{ij}) = \beta_{0j} \qquad (4.5)$$

Note that the usual lowest level error term e_{ij} is not included in equation (4.5). In the binomial distribution the variance of the observed proportion depends only on the population proportion Π. As a consequence, in the model described by equation (4.5) the lowest level variance is determined completely by the predicted value for P_{ij}, therefore it does not enter the model as a separate term.[2] In the models presently analyzed by VARCL and MLn the equation for the lowest level variance is given by:

$$\text{VAR}(P_{ij}) = \sigma^2\,(\Pi_{ij} * (1-\Pi_{ij})) / n_{ij} \qquad (4.6)$$

In equation (4.6) σ^2 is a scale factor. Choosing the binomial distribution in VARCL fixes σ^2 to a default value of 1.00. This means that the binomial model is assumed to hold precisely, and the value 1.00 for the lowest level variance σ^2 is not interpreted. It is possible to fix σ^2 at a value different from 1.00 to model deviations from the binomial distribution, such as overdispersion caused by unmodeled grouping. If that is done, the implication is that equation (4.5) is

[1]We ignore complications arising from different group sizes here, the multilevel software makes all necessary adjustments automatically.

[2]This is similar to the meta-analysis model in section 4.1. In both cases the lowest level variance is known. However, in the meta-analysis model this variance must be supplied, while in the model for proportions it is automatically supplied because it is a function of the estimate P.

extended with an extra error term e_{ij}, the overdispersion factor.[1]

The model in equation (4.5) can be extended to include an explanatory variable X_{ij} (e.g., a variable describing the condition as a mail or face-to-face condition) at the condition level:

$$\text{logit}(P_{ij}) = \beta_{0j} + \beta_{1j} X_{ij} \qquad (4.7)$$

The regression coefficients beta are assumed to vary across studies, and this variation is modeled by the study level variable Z_j in the usual second level regression equations:

$$\beta_{0j} = \gamma_{00} + \gamma_{01} Z_j + u_{0j} \qquad (4.8a)$$
$$\beta_{1j} = \gamma_{10} + \gamma_{11} Z_j + u_{1j} \qquad (4.8b)$$

By substituting (4.8a) and (4.8b) into (4.7) we get the multilevel model:

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{10} X_{ij} + \gamma_{01} Z_j + \gamma_{11} X_{ij} Z_j + u_{0j} + u_{1j} X_{ij} \qquad (4.9)$$

It should be kept in mind that the interpretation of the regression parameters in (4.9) is *not* in terms of the response proportions we want to analyze, but in terms of the underlying variate defined by the logit transformation $\text{logit}(x) = \ln(x/(1-x))$. The logit link function transforms the proportions, which are between 0.00 and 1.00 by definition, into values that range from $-\infty$ to $+\infty$. The logit link is nonlinear, and in effect assumes that near the extremes of 0.00 and 1.00 it becomes more difficult to produce a change in the dependent variable (the proportion). For a quick examination of the analysis results we can simply inspect the regression parameters as calculated by the program. To understand the implications of the regression coefficients for the proportions we are modeling, we must transform their values back to the original scale.[2]

With VARCL, specifying the binomial distribution automatically selects the logit link function. Next, the user has to provide the dependent variable. VARCL

[1]VARCL can fix the extrabinomial variance to some value other than 1. MLn can also estimate the amount of overdispersion and test it for significance.

[2]The inverse function for the logit is $g(x)=\exp(x)/(1+\exp(x))$. We could call this the *expit*.

does not expect a proportion as direct input; the user must provide the variable that contains the number of completed interviews and the variable that contains the so-called `binomial counts', in our example the total number of eligible respondents in the sample.[1]

The `intercept-only' model for our example data is given by the lowest level regression model:

$$\text{logit}(P_{ij}) = \beta_{0j} \tag{4.10}$$

where the random coefficient β_{0j} is modeled by:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{4.11}$$

which leads by substitution to:

$$\text{logit}(P_{ij}) = \gamma_{00} + u_{0j} \tag{4.12}$$

Table 4.2 below presents the results for the 'intercept-only' model:

Table 4.2 Response rates, intercept-only model

intercept (γ_{00}) .72			
	var	sigma	SE(s)
condition level (σ^2)	1.00	1.00	
study level (u_{00})	.57	.75	.08

The intercept γ_{00} is estimated as 0.72. As noted before, this refers to the underlying distribution established by the logistic link function, and *not* to the proportions themselves. To determine the expected proportion, we must use the inverse transformation for the logistic link function, given by g(x)=exp(x)/(1+exp(x)). Using this inverse function we obtain: exp(0.72)/(1+exp(0.72)=2.05/3.05=0.67. Thus, the estimated intercept of 0.72 translates back to an expected proportion of 0.67. This is not precisely equal to the value of 0.69 that we get as the mean proportion

[1]If the dependent variable is dichotomous, the binomial count is given as 1. This models a binomial distribution with 1 trial, which is also known as the Bernouilli distribution.

computed by VARCL just after it has finished reading the data. However, this is as it should be, since we are using a nonlinear link function, and the value for the intercept refers to the intercept of the underlying variate. Transforming that value back to a proportion is *not* the same as computing the intercept for the proportions themselves. Nevertheless, when the proportions are not very close to 1 or 0, the difference is usually rather small.

The value of precisely 1.00 for the variance at the lowest level looks a bit strange. As I explained above, in the binomial distribution (and also in the Poisson and gamma distributions supported by VARCL), the lowest level variance is completely determined when the mean (which in the binomial case is the proportion) is known. Therefore, in these models σ^2 has no useful interpretation; it simply defines the scale for the underlying normal variate. By default σ^2 is fixed at 1.00, which is equivalent to the assumption that the binomial (Poisson, gamma) distribution holds exactly. In some applications the variance of the error distribution turns out to be much larger than expected; there is *overdispersion* (cf. McCullagh & Nelder, 1989; Aitkin et al., 1989). Such overdispersion may be modeled by setting σ^2 to a value larger than 1.00 (in VARCL) or by estimating the amount of overdispersion (in MLn). Since σ^2 is not interpreted, I leave it out of all subsequent tables.

The next model adds the condition level dummy variables X_1 for `telephone-dummy' and $X_2$ for `mail-dummy,' assuming fixed regression slopes. The equation at the lowest (condition) level is:

$$\text{logit}\,(P_{ij}) = \beta_{0j} + \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij}, \qquad (4.13)$$

and at the study level:

$$\beta_{0j} = \gamma_{00} + u_{0j} \qquad (4.14a)$$

$$\beta_{1j} = \gamma_{10} \qquad (4.14b)$$

$$\beta_{2j} = \gamma_{20} \qquad (4.14c)$$

By substituting (4.14a) to (4.14c) into (4.13) we obtain:

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + u_{0j} \qquad (4.15)$$

The results are in Table 4.3:

Table 4.3. Response rates, fixed effect model

	coeff.	SE	p
Intercept	.99		
X_1: tel-dummy	-.25	.02	.00
X_2: mail-dummy	-.62	.03	.00
	var	sigma	SE(s)
study level (u_{00})	.52	.72	.08

The intercept represents the condition in which both explanatory variables X_1 and X_2 are zero. When X_1 (telephone-dummy)=0 and X_2 (mail-dummy)=0, we have the face-to-face condition. Thus, the value for the intercept Table 4.3 estimates the expected response in the face-to-face condition, and the expected response in this condition Y_{ftf} equals 0.99. The large negative values for the slope coefficients for the telephone and mail dummy-variables indicate that in these conditions the expected response is much lower. To find out how much lower, we must use the regression equation to predict the response in the three conditions, and transform these values (which refer to the underlying variate) back to proportions. For the telephone condition, that is coded by X_1=1 and X_2=0, the regression equation reads: Y=0.99-0.25=0.74, and for the mail condition, that is coded by X_1=0 and X_2=1, it reads Y=0.99-0.62=0.37. The predicted values for Y in the three conditions again refer to the underlying variable, which has to be transformed back to proportions for interpretation (using the inverse transformation for the logit function given earlier). Translating back to proportions gives us predicted proportions of 0.73 in the face-to-face condition, 0.68 in the telephone condition, and 0.59 in the mail condition.

The intercept variation σ_{00} on the study level is obviously significant, and we may attempt to explain it by the known differences between the studies. In our

example data we have three study level explanatory variables: year of publication, salience of questionnaire topic, and research quality. Since not all studies compare all three data collection methods, it is quite possible that study level variables also explain between condition variance. For instance, if older studies tend to have a higher response rate, and the telephone method is only included in the more recent studies (telephone interviewing is, after all, a relatively new method), the telephone condition may seem to be characterized by low response rates. In that case, however, after correcting for the year of publication, the telephone response rates should look better. We cannot inspect the condition level variance to see if the higher level variables explain condition level variability, because the condition level variance is fixed at 1.00, but we can scrutinize the regression coefficients for the two dummy variables coding the telephone and mail condition to see whether there is a substantial change when we go from one model to the next.

Of the three study level variables, only saliency turns out to make makes a significant contribution to the regression equation. The equations for the model including the study level explanatory variable saliency (Z_1) are:

At the condition (lowest) level:

$$\text{logit}(P_{ij}) = ß_{0j} + ß_{1j} X_{1ij} + ß_{2j} X_{2ij} \qquad (4.16)$$

and at the study level:

$$ß_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + u_{0j} \qquad (4.17a)$$
$$ß_{1j} = \gamma_{10} \qquad (4.17b)$$
$$ß_{2j} = \gamma_{20} \qquad (4.17c)$$

By substituting (4.17a to (4.17c) into (4.16) we obtain:

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_{1j} + u_{0j} \qquad (4.18)$$

The results are in Table 4 4:

Table 4.4. Response rates, fixed effect model, including saliency

	coeff.	SE	p
Intercept	1.97		
X_1: tel-dummy	-.26	.02	.00

X_2: mail-dummy	-.62	.03	.00
Z_1: saliency	-.50	.14	.00
	var	sigma	SE(s)
study level (u_{00})	.40	.63	.07

Compared to the earlier results, the regression coefficients are about the same, but the value for the intercept is different. This is not informative, because the intercept almost always changes when other variables are added to or deleted from the regression equation. In our case, the shift of the intercept value is caused by including the study level explanatory variable `saliency' in the model. Saliency is coded as: 1=very salient, 2=somewhat salient, and 3=not salient. The coded values for `saliency' do not include the value 0. Since the value of 1.97 for the intercept refers to the situation where all explanatory variables have the value zero, this refers to a face to face survey where the saliency is equal to zero, meaning it is extremely high, in fact beyond the saliency range in the set of studies under review.

Until now, we have treated the two dummy variables as fixed. One could argue that it doesn't make sense to model them as random, since the dummy variables are simple dichotomies that code for our three experimental conditions. The experimental conditions are under control of the investigator, and there is no reason to expect their effect to vary from one experiment to another. But some more thought leads to the conclusion that the situation is more complicated than it seems. If we conduct a series of experiments, we would expect identical results only if the research subjects were all sampled from exactly the same population, and if the operations that define the experimental conditions were all carried out in exactly the same way. In the present case, both assumptions are questionable. In fact, some studies have sampled the general population, while others sample special populations such as college students. Similarly, although most articles give only a very short description of the procedures that were actually used to implement the data collection methods, it is highly likely that they were not all identical. As a consequence, even if we don't know all the details about the populations sampled and the procedures used, we may expect much variation

between the conditions in the way they actually were implemented. This should result in random regression coefficients in our model. Thus, we analyze a model in which the slope coefficients for the dummy variables X_1 (telephone-dummy) and X_2 (mail-dummy) are assumed to be random across studies.

At the condition (lowest) level we have:

$$\text{logit}(P_{ij}) = ß_{0j} + ß_{1j} X_{1ij} + ß_{2j} X_{2ij} \qquad (4.19)$$

and at the study level:

$$ß_{0j} = \gamma_{00} + \gamma_{01} Z_{1j} + u_{0j} \qquad (4.20a)$$
$$ß_{1j} = \gamma_{10} + u_{1j} \qquad (4.20b)$$
$$ß_{2j} = \gamma_{20} + u_{2j} \qquad (4.20c)$$

By substituting (4.20a) to (4.20c) into (4.19) we obtain:

$$\text{logit}(P_{ij}) = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_{1j} +$$
$$+ u_{0j} + u_{1j} X_{1ij} + u_{2j} X_{2ij} \qquad (4.21)$$

Table 4.5 below presents the estimates for the model of (4.21):

Table 4.5 Response rates, random coefficient model

	coeff	SE	p
Intercept	2.04		
X_1: tel-dummy -.23	.11	.04	
X_2: mail-dummy	-.54	.12	.00
Z_1: saliency	-.56	.12	.00
study level	var	sigma	SE(s)
intercept (u_{00}) .33	.57	.07	
telephone (u_{11}) .39	.63	.08	
mail (u_{22})	.31	.56	.10

Indeed, the variance of the regression slopes for the two dummy variables is large and highly significant. The last logical step would be to introduce interaction variables to model the random coefficients. In our example data, it turns out that none of the available interaction variables explains the random variation of the regression coefficients (none of the regression slopes for the cross-level interactions even approached significance). It is likely that the variance of the regression slopes over studies is the result of uncountable variations in the way the different data collection methods were actually implemented. Unfortunately, the articles reviewed in the meta-analysis do not give all the necessary details, and therefore it is impossible to define and code additional explanatory study level variables to model them, so they show up in the random variation of the regression coefficients.

As I noted above, the regression coefficients have to be interpreted in terms of the underlying variate. Also, the logit transformation implies that raising the response becomes more difficult as we approach the limit of 1.00. To show what this means, I present the predicted response for the three methods as logits (in parentheses) and proportions in the next table, both for a very salient (saliency=1) and a non-salient (saliency=3) questionnaire topic. (To compute these numbers we must fill in the regression equation implied by Table 4.5 and use the inverse logit transformation given earlier to transform the predicted logits back to proportions.)

Table 4.6 Response rates for the three methods, based on Table 4.5

Topic	Face-to-face		Telephone		Mail	
Not salient	(.37)	.59 (.14)		.54 (-.17)		.46
Very salient		(1.49) .82		(1.26) .79		(.95) .72

When the topic is very salient, the response rate is generally high, and in this condition the advantage of the face-to-face survey is smaller than with a non-salient topic, where the face-to-face situation apparently is a much better method to persuade potential respondents to cooperate.

The random coefficient model leads to another interesting conclusion. In general, telephone surveys obtain a lower response rate than face-to-face surveys. On the underlying normal scale, the regression coefficient for `telephone' is -0.23. However, this regression coefficient has a large variance across studies: σ_{11}=0.39. The corresponding standard deviation is 0.63. Using the standard normal distribution, we can calculate that in 36% of similarly conducted studies this regression coefficient is actually larger than zero! Since in the binomial distribution the distribution of the random regression coefficients is probably skewed, we should not interpret this 36% as an exact prediction of what would happen in new replications. Still, it is instructive to see that, even if there is little doubt that *on the average* the telephone interview has a lower response rate than the face-to-face interview, there is still an appreciable chance that *in a specific study* we find the opposite relation.

5. Structural Models for Multilevel Data

The models described in the previous chapters are all basically multilevel variants of the conventional multiple regression model. This is not as restrictive as it may seem, since the multiple regression model is very flexible and can be used in many different applications (for detailed examples see Cohen & Cohen, 1983). Still, there are models that cannot be analyzed with multiple regression, notably factor analysis and path analysis models.

A general approach that encompasses both factor and path analysis is covariance structure analysis. Covariance structure models (sometimes simply but inaccurately denoted as `Lisrel-models') can be viewed as a combination of a path model and a factor model. The path model, which is often called the structural model, specifies causal and predictive relationships between variables. These variables may be observed variables and/or latent factors. The factor model, which is often called the measurement model, specifies how the latent factors are measured by the observed variables. The name `covariance structure analysis' (CSA) derives from the usual practice of using the mathematical model to describe the covariance matrix of the observed variables. Other names for this model are structural equations models (SEM), structural models or path models with latent variables.

Structural models for multilevel data have been elaborated, among others, by Goldstein and McDonald (Goldstein & McDonald, 1988; McDonald & Goldstein, 1989), Muthén and Satorra (Muthén, 1989; Muthén & Satorra, 1989) and Longford and Muthén (Longford & Muthén, 1992). Simple introductions are given by Muthén (1994) and McDonald (1994), and the likelihood equations are given in the literature cited above. This chapter focusses on a simplification proposed by Muthén (1989), which makes it possible to use existing software (such as Lisrel, Eqs, Liscomp) for the covariance structure analysis of multilevel data.

5.1 The Decomposition Model for a Hierarchical

Population

Suppose we have data from N individuals, divided into G groups. The individual data are collected in a p-variate vector $\mathbf{Y}_{ig}$ (subscript i for individuals, i=1..N; subscript g for groups, g=1..G). Cronbach and Webb (1975) propose to decompose the individual data $\mathbf{Y}_{ig}$ into a between groups component $\mathbf{Y}_B = {}_{.g}$ and a within groups component $\mathbf{Y}_W = \mathbf{Y}_{ig} - {}_{.g}$. In other words, for each individual we replace the observed *Total* score $\mathbf{Y}_T = \mathbf{Y}_{ig}$ by its components: the group component $\mathbf{Y}_B$ (the disaggregated group mean) and the individual component $\mathbf{Y}_W$ (the individual deviation from the group mean.) These two components have the attractive property that they are orthogonal and additive:

$$\mathbf{y}_T = \mathbf{y}_B + \mathbf{y}_W \tag{5.1}$$

This decomposition can be used to compute a between groups covariance matrix $\mathbf{S}_B$ (the covariance matrix of the disaggregated group means $\mathbf{Y}_B$) and a within groups covariance matrix $\mathbf{S}_W$ (the covariance matrix of the individual deviations from the group means $\mathbf{Y}_W$). The covariance matrices are also orthogonal and additive:

$$\mathbf{S}_T = \mathbf{S}_B + \mathbf{S}_W \tag{5.2}$$

Härnqvist (1978) proposes to apply exploratory factor analysis to $\mathbf{S}_B$ and $\mathbf{S}_W$. However, the decomposition in equation (5.1) is a decomposition of the sample data. Covariance structure models are confirmatory models for a population, and to apply them to multilevel data we have to examine what the consequences are of the decomposition proposed by Cronbach and Webb in the population.

Multilevel structural models assume that we have a population of individuals that are divided into groups. If we decompose the population data we have, for the population covariance matrices:

$$\mathbf{\Sigma}_T = \mathbf{\Sigma}_B + \mathbf{\Sigma}_W \tag{5.3}$$

Covariance structure modeling assumes that the population covariance matrices $\mathbf{\Sigma}_B$ and $\mathbf{\Sigma}_W$ can be described by separate models for the between groups and within

groups structure. Unfortunately, we cannot simply use $\mathbf{S}_B$ as an estimate of $\mathbf{\Sigma}_B$, and $\mathbf{S}_W$ for $\mathbf{\Sigma}_W$. The situation is a bit more complicated.

Muthén (1989) shows that an unbiased estimate of the population within groups covariance matrix $\mathbf{\Sigma}_W$ is given by the pooled within groups covariance matrix $\mathbf{S}_{PW}$, calculated in the sample by:

$$\mathbf{S}_{PW} = \frac{\sum_{G}\sum_{n} (Y_{ig} - Y_{.g})(Y_{ig} - Y_{.g})'}{N - G} \qquad (5.4)$$

Equation (5.4) corresponds to the conventional equation for the covariance matrix of the individual deviation scores, with N-G in the denominator instead of the usual N-1.

Since the pooled within groups covariance matrix $\mathbf{S}_{PW}$ is an unbiased estimate of the population within groups covariance matrix $\mathbf{\Sigma}_W$, we can estimate the population within group structure by constructing and testing a model for $\mathbf{S}_{PW}$.

The between groups covariance matrix for the disaggregated group means $\mathbf{S}_B$, calculated in the sample is given by:

$$\mathbf{S}_B = \frac{\sum_{G} n_g (Y_{..} - Y_{.g})(Y_{..} - Y_{.g})'}{G} \qquad (5.5)$$

Unfortunately, the sample between groups covariance matrix $\mathbf{S}_B$ is not a simple estimator of the population between groups covariance matrix $\mathbf{\Sigma}_B$. Instead, $\mathbf{S}_B$ is an estimator of the sum of two matrices:

$$\mathbf{S}_B = \hat{\mathbf{\Sigma}}_W + c\mathbf{\Sigma}_B \qquad (5.6)$$

where c is a scaling factor based on the group size.

Thus, if we want to model the between groups structure, we cannot simply construct and test a model for $\mathbf{S}_B$, because $\mathbf{S}_B$ estimates a combination of $\mathbf{\Sigma}_W$ and $\mathbf{\Sigma}_B$. Instead, we have to specify for $\mathbf{S}_B$ two models: one for the within groups structure and one for the between groups structure. Muthén (1989) proposes to use the multigroup option of conventional covariance structure software to analyze these models simultaneously. The procedure is that we specify two groups, with covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_B$ (based on N-G and G observations). The model for $\mathbf{\Sigma}_W$ must be specified for both $\mathbf{S}_{PW}$ and $\mathbf{S}_B$, with equality restrictions between both `groups' to guarantee that we are indeed estimating the same model in both covariance matrices, and the model for $\mathbf{\Sigma}_B$ is specified for $\mathbf{S}_B$, with the scale factor c built into the model.

The reasoning given above applies only in the so-called *balanced* case, that is, if all groups have the same group. In the balanced case, the scale factor c is equal to the common group size n. In the unbalanced case, where the group sizes differ, using conventional software requires a complicated modeling scheme that creates a different `group' for each set of groups with the same group size. In many cases this is not practical. As a solution, Muthén (1989, 1990) proposes to simply proceed as if the group sizes were equal, and calculate the scaling factor as a combination of the observed group sizes given by:

$$C = \frac{N^2 - \sum_{G} n^2_g}{N\,(G-1)} \qquad (5.7).$$

This solution, which McDonald (1994) calls a *pseudobalanced* solution, is not a full likelihood solution. However, Muthén (1990) shows that $\mathbf{S}_B$, as calculated in equation (5.5), is a *consistent* estimator of $\mathbf{\Sigma}_B$. This means that with large samples (of both individuals *and* groups!) $\mathbf{S}_B$ generally becomes a close estimate of $\mathbf{\Sigma}_B$. Since $\mathbf{S}_B$ is not a maximum likelihood estimator, the analysis produces only approximate parameter estimates and standard errors. However, when the group sizes are not extremely different, the pseudobalanced estimates may be close enough to the full maximum likelihood estimates to be useful in their own right. Comparisons of

pseudobalanced estimates with full maximum likelihood estimates or with known population values have been made by Muthén (1990, 1994), Hox (1993), and McDonald (1994). Their main conclusion is that the pseudobalanced estimates are fairly accurate and useful for a variety of multilevel problems.

The multilevel part of the covariance structure model outlined above is simpler than that of the multilevel regression model. It is comparable to the multilevel regression model with random variation of the intercepts. There is no provision for randomly varying slopes (factor loadings and path coefficients). Although it would be possible to include cross-level interactions, introducing interaction variables of any kind in covariance structure models is neither simple nor elegant (cf. Bollen, 1989). An interesting approach would be to allow for different within groups covariance matrices in different subsamples.

5.2 An Example of a Multilevel Factor Analysis

The example data are taken from Van Peet (1992). They are the scores on six intelligence measures of 187 children from 37 families. The six intelligence measures are: word list, cards, matrices, figures, animals, and occupations. The data have a multilevel structure, with children nested within families. Assuming that intelligence is strongly influenced by shared genetic and environmental influences in the families, we may expect rather strong between family effects.

To begin, the individual scores on the six measures are decomposed into disaggregated group means and individual deviations from the group means (Cronbach & Webb, 1975, cf. section 5.1). Table 5.1 shows the means and variances of the scores, and the Intra Class Correlation (ICC), which is an estimate of the proportion of between family variance.[1]

Table 5.1. Means, variances and ICC for family data

Measure	Total Mean	Total Var.	Family Var.	Individual Var.	ICC

[1]The ICC can be estimated by analysis of variance procedures (Hays, 1994), or from the intecept-only model using a multilevel approach. Here, it is estimated from the pooled within groups and between groups variances.

Word list	29.80	15.21	7.48	7.73	.37
Cards	32.68	28.47	13.65	14.82	.35
Matrices	31.73	16.38	5.24	11.14	.15
Figures	27.11	21.23	6.84	14.38	.16
Animals	28.65	22.82	8.46	14.36	.22
Occupat.	28.28	21.42	9.11	12.31	.28

The results in Table 5.1 suggest that there are indeed sizeable family effects. To analyze the factor structure of the six measures on the individual and family level, we compute the pooled within family covariance matrix $\mathbf{S}_{PW}$ and the between family covariance matrix $\mathbf{S}_{B}$.

Typically, there are many more individuals than groups, and hence the number of observations for the pooled within groups covariance matrix (N-G) is much larger than the number of observations for the between groups covariance matrix (G). In this case, the number of observations on the individual level is 187-37=150, while on the family level it is 37. Thus, it makes sense to start on the individual level by constructing a model for $\mathbf{S}_{PW}$.

An exploratory factor analysis on $\mathbf{S}_{PW}$ suggests two factors, with the first three measures loading on the first factor, and the last three measures on the last. A confirmatory factor analysis on $\mathbf{S}_{PW}$ confirms this model: χ^2=7.21, df=8, p=.51. A model with just one general factor is rejected: χ^2=44.87, df=9, p=.00. Figure 5.1 on the next page presents the conventional graphic representation of the individual level (within families) model.

The next step is the specification of a family model. For this, we must analyze the matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_{B}$ simultaneously with the multigroup procedure. First we specify the individual model for both `groups' using equality restrictions across both groups for all parameters. Next, we must specify an additional family model for $\mathbf{S}_{B}$.

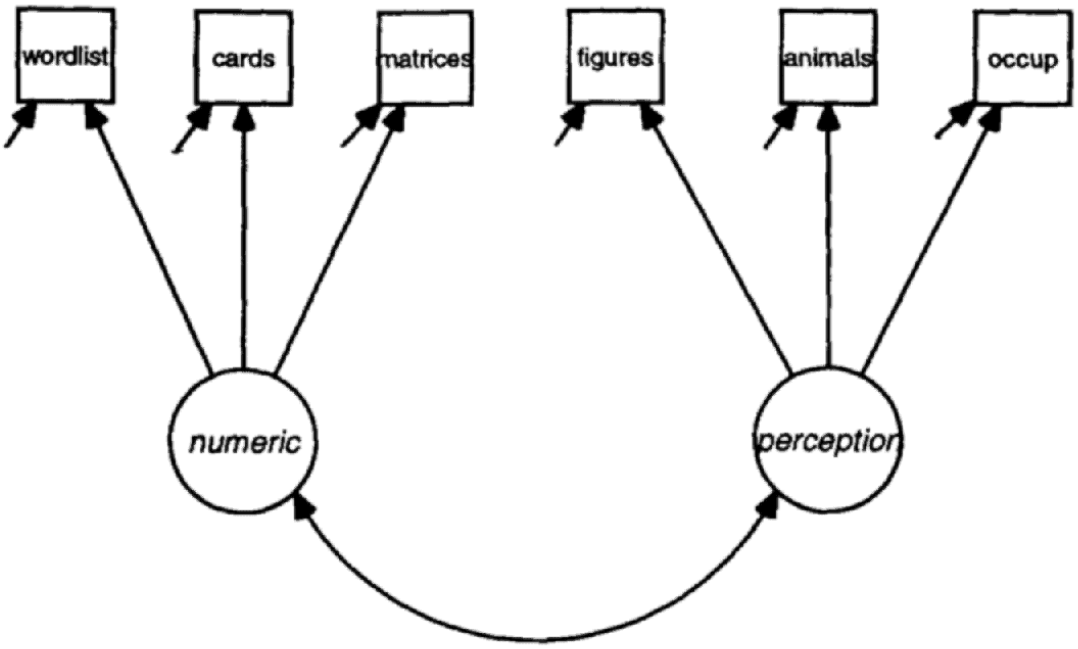


Figure 5.1 Within families model for Van Peet data.

We start by estimating some `benchmark' models, to test whether there is any between family structure at all. The most simple model is the null model; this simply omits the specification of a family level model. If the null model holds, there is no family level structure; all covariances in $\mathbf{S}_B$ are the result of sampling individual variation. As a result, we may as well continue our analyses using simple single level analysis methods. The next model is the independence model; this specifies only variances on the family level, but no covariances. A graphical representation of the independence model for $\mathbf{S}_B$ is given in Figure 5.2 on the next page.

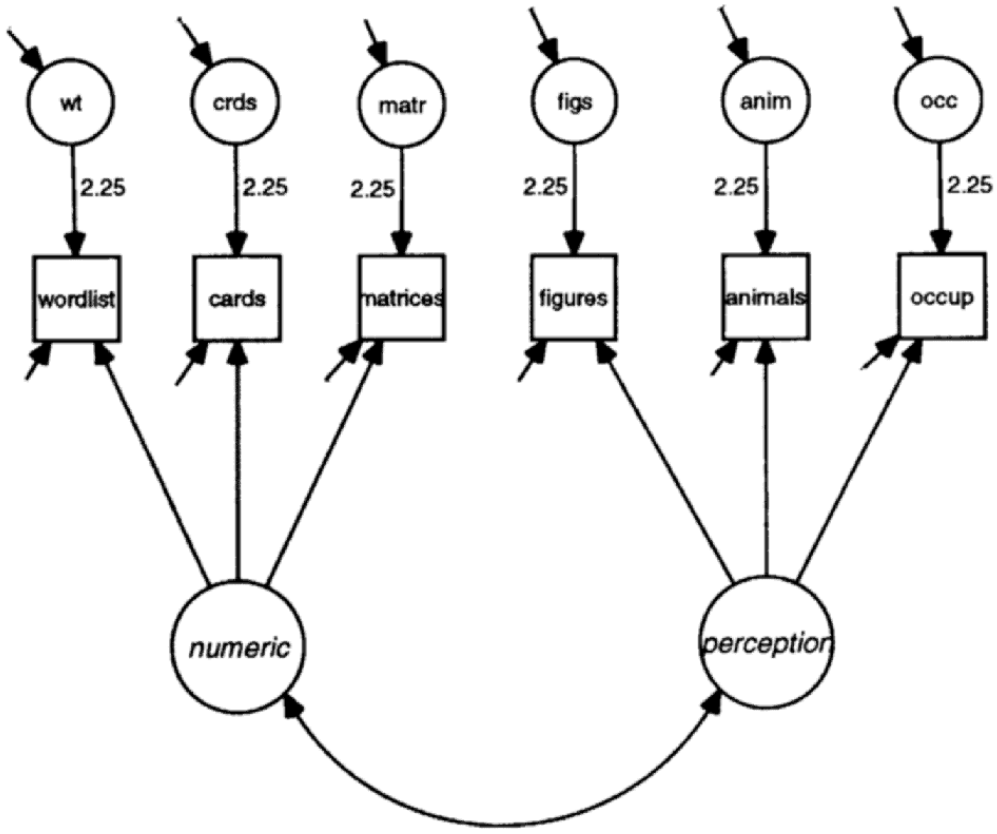


Figure 5.2 Within + Between (independence) families model for Van Peet data.

Note that in Figure 5.2 I have fixed the loadings for the family level variables (the six `factors' in the circles going from `wl' to `occ') not to one, as is usual, but to 2.25, which is the square root of the scale factor. This is to transform the family level variables to their proper scale. Since this is a fixed value, it has no influence on the global fit of the model, but it is necessary for a correct interpretation.

If the independence model holds, there is family level variance, but no substantively interesting covariance structure. Nevertheless, in this case it is still useful to apply multilevel analysis, because this produces unbiased estimates of the individual model parameters. If the independence model is rejected, there is some kind of covariance structure on the family level. To examine the best possible fit given the individual level model, we can estimate the maximal model; this fits a full covariance matrix to the family level observations. This places no restrictions

on the family model.[1] Table 5.2 shows the results of estimating these models:

Table 5.2 Comparison of family level benchmark models

Family model	Chi-square	df	p
Null model	125.41	29	.00
Independence model	52.47	23	.00
Maximum model	7.21	8	.51

Both the null model and the independence model are rejected. Next, we specify for the family level the same two models we have used for the individual level. Again, the two factor model fits well. However, on the family level a one factor model fits almost as well, as Table 5.3 shows:

Table 5.3 Comparison of family level factor models

Family model	Chi-square	df	p
One factor	21.28	17	.21
Two factors	20.06	16	.22

The principle of using the simplest model that fits well leads to acceptance of the one factor model on the family level, see Figure 5.3 on the next page. The factor loadings (standardized to a common metric) are in Table 5.4:

[1]We can specify the maximal model for the between structure, and then explore the within model simultaneously in both $\mathbf{S}_{PW}$ and $\mathbf{S}_{B}$. However, since $\mathbf{S}_{PW}$ is generally based on many more observations than $\mathbf{S}_{B}$, not much information is lost by only analyzing $\mathbf{S}_{PW}$, while in the latter case the setups are much simpler and need less computing time.

Table 5.4 Individual and family model, standardized factor loadings

		Individual		Family
		I	II	I
Word list		.30*	-	.84*
Cards	.52	-	.78	
Matrices		.70	-	1.02
Figures		-	.30	.58
Animals		-	.70	.86
Occupations		-	.48	.33^{n}

Correlation between individual factors: 0.22^{ns}; * = fixed; ns = not significant

Table 5.4 suggests an interpretation that on the family level, where the effects of the shared genetic and environmental influences are visible, one general (*g*) factor is sufficient to explain the covariances between the intelligence measures. On the individual level, where the effects of individual idiosyncratic influences are visible, we need two factors. The first factor could be interpreted as `reasoning,' and the second as `fluency.' These results could be fitted into Cattel's (1971) theory of fluid and crystallized intelligence, which states that as a result of individual factors (education, physical and social environment) the general g-factor `crystallizes' into specific individual competencies.

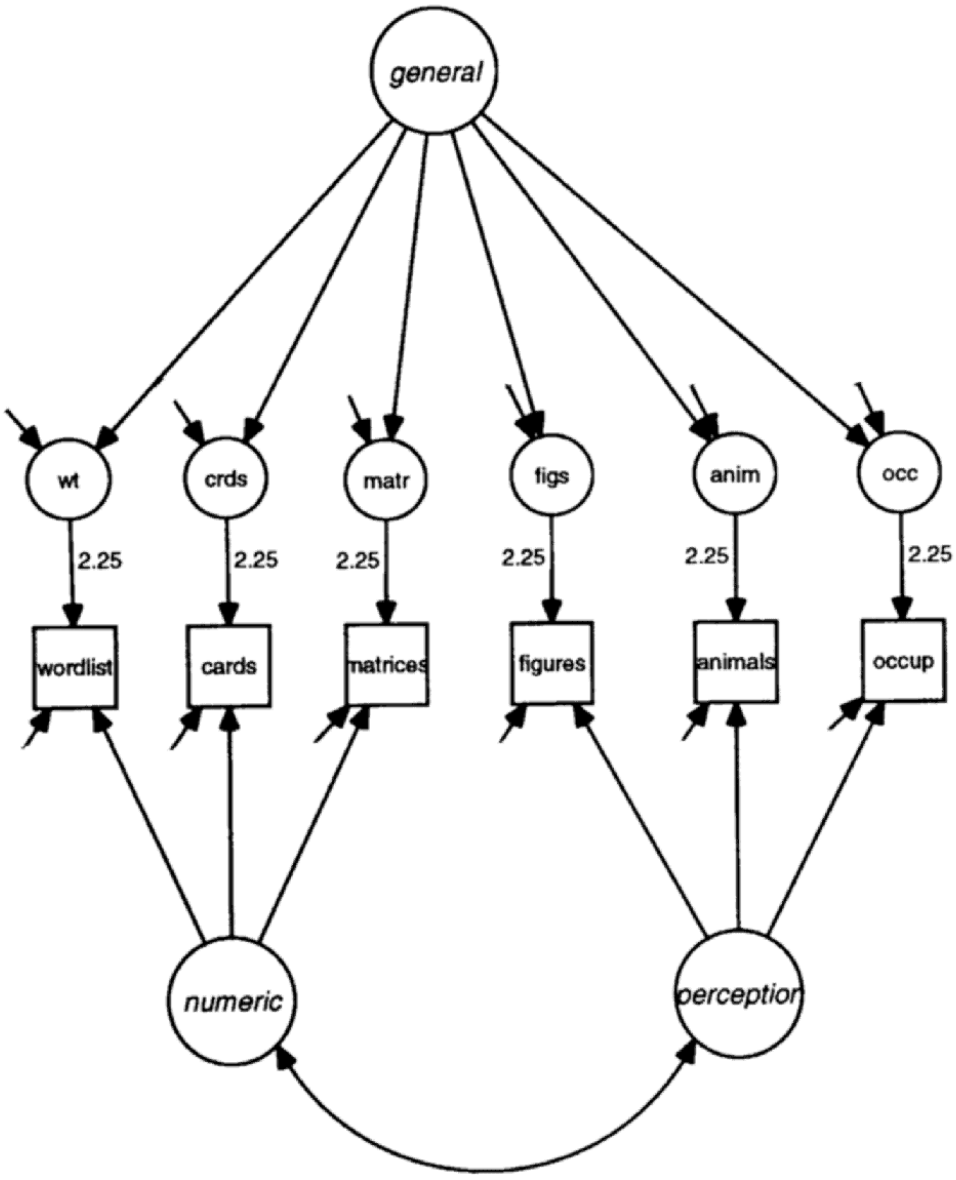


Figure 5.3 Within + Between families model for Van Peet data.

5.3 An Example of a Multilevel Path Analysis

The data for this example are from a study by Schijf and Dronkers (1991). They analyzed data from 1379 pupils in 58 schools.[1] We have the following pupil level variables: father's occupational status `focc,' father's education `feduc,' mother's education `meduc,' family size `fsize,' sex `sex,' how many times a class had been repeated in the past `repeat,' result of GALO school test `GALO,' and teacher's advice about secondary education `advice.' On the school level we have only one variable: the school's denomination `denom.' Denomination is coded 1=protestant, 2=nondenominational, 3=catholic (categories based on optimal scaling). The research question is whether the school's denomination affects the GALO score and the teacher's advice, after the other individual variables have been accounted for.

We can use a sequence of multilevel regression models to answer this question. The advantage of a path model is that we can specify one model that describes all hypothesized relations between independent, intervening, and dependent variables. However, we have multilevel data, with one variable on the school level, so we must use a multilevel model to analyze these data.

A multilevel path model uses the same approach outlined above for the multilevel factor analysis. We decompose the individual variables into disaggregated group means and individual deviations from the group means, and calculate the pooled within groups covariance matrix $\mathbf{S}_{PW}$ and the between groups covariance matrix $\mathbf{S}_{B}$. Next, we construct models for $\mathbf{\Sigma}_{W}$ and $\mathbf{\Sigma}_{B}$, and use the multigroup approach as illustrated above.

With multilevel path models we will often have the complication that we have pure group level variables (*global* variables in the terminology of chapter one). In our example, we have the global variable `denomination.' This variable does not exist on the individual level. We can of course disaggregate `denomination' to the individual level. However, this disaggregated variable is constant within each school, and as a result the variance and covariances with the individual deviation

[1]The data were collected in 1971. The example uses only those pupils with complete data on all variables.

scores are all zero. This problem can be solved by viewing the school level variable `denomination' as a variable that is systematically missing in the pupil level data. Bollen (1989) and Jöreskog and Sörbom (1989) describe how systematically missing variables can be handled in Lisrel.[1]

Basically, the trick is that the variable `denomination' is included in the (school level) between schools covariance matrix in the usual way. In the (pupil level) within schools covariance matrix, we include `denomination' as a variable with a variance equal to one and all covariances with other variables equal to zero. In the within school models, there are no paths pointing to or from this observed variable. As a consequence, we estimate for this variable only a residual error variance of 1.00. Thus, inclusion of this variable has no influence on the within school estimates or the overall chi-square. There is only one problem; Lisrel assumes that this variable represents a real observed variable, and will include it when it enumerates the degrees of freedom for the within schools model. As a result, the df and p-values (and most fit-indices) in the Lisrel output are incorrect, which must be corrected by hand (cf. Jöreskog & Sörbom, 1989). Some software (notably Bentler's Eqs, cf. Bentler, 1993), can handle multigroup models with different numbers of observed variables in the various groups, which makes this kind of modeling much simpler.

After calculating the pooled within and between groups covariance matrices, the first step in modeling the Schijf/Dronkers data is again to construct a within schools (pupil level) model. I used earlier analyses by Schijf and Dronkers (Dronkers & Schijf, 1986; Schijf & Dronkers, 1990, 1991) to arrive at the following pupil level model shown on the next page.

The pupil level path model has one latent factor `SES' measured by the observed variables `focc,' `fedu' and `medu.' An analysis of the pooled within schools matrix S_{PW} only with this model gives a chi-square of 32.8, with df=15 and p=0.01. The goodness-of-fit indices are: GFI=.99 and AGFI=.99. Given the large sample size (the pupil level data have 1379-58=1321 independent observations,) the high goodness-of-fit and the absence of large modification indices in the Lisrel

[1]Bollen gives a more detailed description of the model, but Jöreskog and Sörbom's account is to be preferred when a model must be specified, because they use some Lisrel features that were not yet available when Bollen wrote his book.

output, I decide to accept this model.

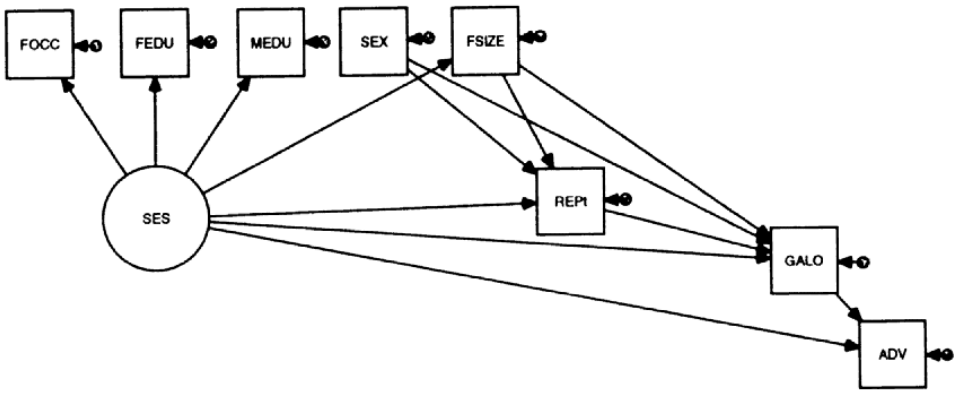


Figure 5.4: Pupil level path model for GALO Data.

Inspection of the intra class correlations tells us that the school level variance for the pupil variable `sex' is zero. This simply means that the proportion of girls is almost the same in all schools. Therefore, this variable is eliminated from the school level covariance matrix by treating it as a variable with systematically missing values, in the same way the school level variable `denomination' is handled on the individual level.[1]

The next step is again specifying a school level model for $\mathbf{S}_B$. First we specify the pupil level model constructed earlier for both $\mathbf{S}_{PW}$ and $\mathbf{S}_B$, with equality restrictions across the two `groups' for all corresponding parameters. We start the analysis of the between groups matrix $\mathbf{S}_B$ by specifying three benchmark models: the null model, the independence model, and the maximum model. This produces the following results:

[1]Since sex is eliminated from the school level on empirical grounds, rather than because of the design of the data, it is not necessary to correct the degrees of freedom.

Table 5.5 School level benchmark models

Model	Chi-square	df	p	
Null	623	51	.00	
Independence	392	44	.00	
Maximum	39	23	.02	

Table 5.5. shows that the null and independence model are rejected, there is some kind of school level covariance structure. The maximum model specifies a saturated model for $\mathbf{S}_B$, meaning that it produces an estimate of the full covariance matrix $\mathbf{\Sigma}_B$.[1] Inspection of this covariance matrix reveals that on the school level the three SES indicators `focc,' `feduc' and `meduc' have extremely high correlations (all intercorrelations are larger than .98). A school-level factor model for these three indicators does not converge, and is therefore replaced by a component model. Thus, on the school level we have a path model without latent variables (other than having latent variables to represent the school level).

The school level model resembles the individual level model, but with fewer significant paths. The fit of the model is acceptable (chi-square=58, df=40, p=.03). The school level variable `denomination' turns out to have an effect on only one variable, the GALO test score. The graphical representation of this final model (chi-squared=64, df=47, p=.05) is given in Figure 5.5 on the next page.

In the multilevel regression analyses presented by Schijf and Dronkers (1991) denomination had a significant effect on both the teachers' advice and on the GALO test score. The path model presented here shows that the main influence is through the GALO test score; the different advice given by teachers in schools of different denominations are apparently the result of differences in GALO test scores between such schools. This is precisely the kind of result that a sequence of regression analyses cannot show.

[1]The null and maximum model can be used to define goodness-of-fit measures. Since the maximum model has a p-value of .02, it is likely that all school level models will be significant. In such situations, goodness-of-fit indices are a useful alternative to significance testing.

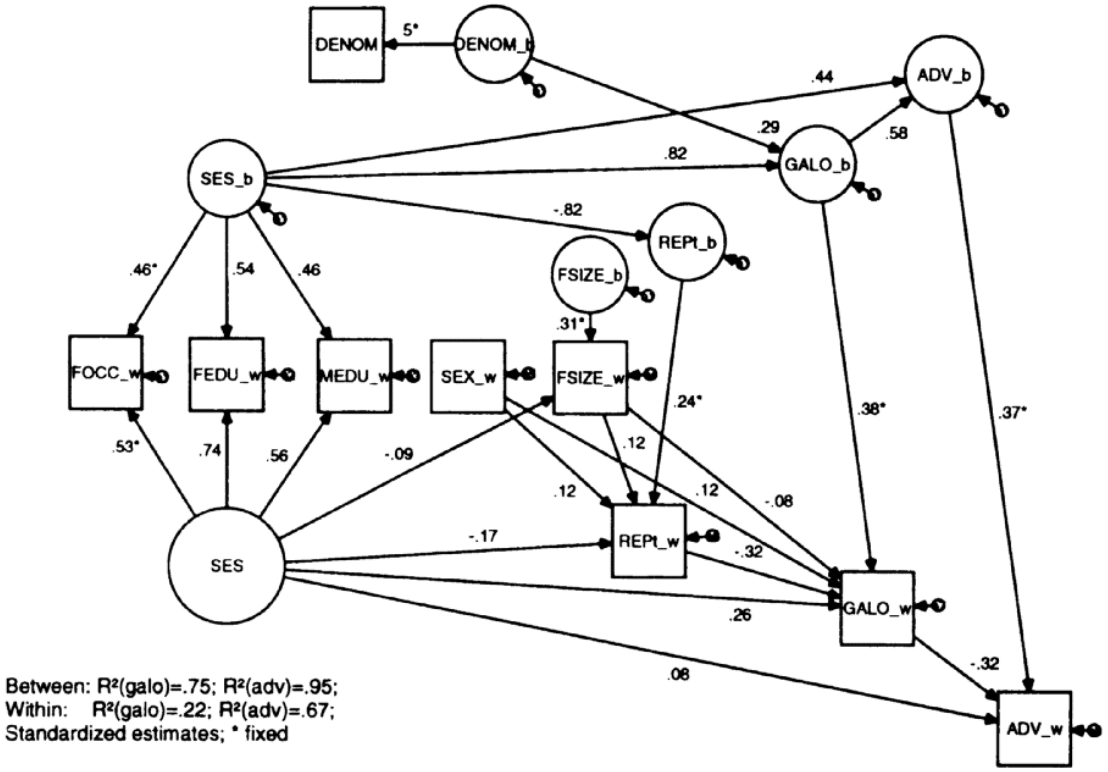


Figure 5.5 Final path model for GALO data, with estimates

Figure 5.5 also shows that `SES' has a school level effect on the variables `repeat,' `GALO,' and `advice.' The interpretation is *not* that some schools simply happen to have more high SES pupils and therefore perform better; sampling differences between schools in SES composition are accounted for in the pupil model that is also fitted for the school level covariances. Instead, the substantive interpretation of the school level results must be in terms of some kind of contextual or systematic selection effect. It appears that the concentration of high or low SES pupils has its own effect on the school career variables.

5.4 Some Implementation Details

For multilevel factor and path analyses we must first compute the within and between groups covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_{B}$, and the scale factor *c*. Implementation of the pseudobalanced model in a conventional covariance structure analysis program such as Lisrel, Liscomp or Eqs is complex. For example, let us examine the Lisrel setup for the factor analysis of the Van Peet data in section 5.2.

The Van Peet data consist of six test-scores for 187 children from 37 families. In the analysis, we treat the covariance matrices $\mathbf{S}_{PW}$ and $\mathbf{S}_{B}$ as if they came from two different groups, with 150 and 37 observations. It is convenient to let $\mathbf{S}_{PW}$ be the first group. The two-factor model for $\mathbf{S}_{PW}$ is specified as usual, with the factor loadings in the Lisrel matrix Lambda-Y, the covariances between the two factors in the matrix Psi, and the residual error variances in the vector Theta-Epsilon.

To specify the complete model in Figure 5.3, we must set the number of factors at *nine*: the two regular factors for $\mathbf{S}_{PW}$, and for $\mathbf{S}_{B}$ the six factors that correspond to the six variables at the between families level, plus the one factor that is sufficient to explain the between families covariation. The last seven factors are not used in the model for $\mathbf{S}_{PW}$, and as a result Lisrel sets all their loadings, covariances etcetera to zero. In the model for $\mathbf{S}_{B}$ they are used. The six factors that correspond to the six variables at the between families level have fixed loadings on their corresponding variable with loading equal to $\sqrt{c}$, the square root of the scaling factor *c*. The one between families factor that explains the between families covariation has loadings on all six between families variables. These are represented as loadings from a factor on six other factors, which in Lisrel is specified by estimating the corresponding elements in the regression matrix called Beta. The residual variances for the between families model appear in the diagonal of the factor covariance matrix Psi.

Since the between covariance matrix is orthogonal to the within covariance matrix, the covariances between the within factors and all between factors must be specified as zero. Also, the covariances among the six factors representing the between families variables and the one factor used to explain them must be set at zero.

If we have a group variable, which has no variation on the within level, things get even more complicated. For this group variable, we have in the pooled within covariance matrix a row and column with all zeros and a one on the diagonal. To model this variable, we must specify on the within groups level zero factor loadings, and leave the residual error variance in theta-epsilon free, which will of course be estimated as one. Since Lisrel will count this extra variable as an ordinary observed variable, it will calculate the number of degrees of freedom incorrectly, something which must then be adjusted by hand. All this leads to a complicated Lisrel model. One result of these complications is that Lisrel diagnoses the model as inadmissible, so the `admissibility check' must be set off. Another complication is that Lisrel often cannot automatically find good starting values for such models, and needs much more iterations than for the more usual models.

Software that specifies the model in the form of equations, such as Eqs, is somewhat easier to use. However, this software generally attempts to simplify the modeling process by automatically assuming covariances between latent variables, while for the between model most of these must be fixed at zero.[1]

For the computation of the pooled within and scaled between groups matrix it is useful to employ a special preprocessor. Muthén has made available the program BW (Muthén, 1989; Nelson & Muthén, 1991) that computes all statistics needed for either the full maximum likelihood solution or the pseudobalanced solution. Included with this book is the simpler program SPLIT2 that only provides statistics for the pseudobalanced solution.

[1]Lisrel8 comes with the simplified command language Simplis that is also equation-based, and sidesteps all reference to the usual Lisrel matrices. Unfortunately, Simplis cannot handle the more complex multilevel specifications. In general, it will not succeed in translating these correctly into the usual matrix oriented Lisrel language.

Appendix

Aggregating and Disaggregating in SPSS

A common procedure in multilevel analysis is to aggregate individual level variables to higher levels. On the higher level, the scheme of page 2 calls such a variable an analytical or a structural variable. In most cases, aggregation is used to attach to higher level units (e.g., groups, classes, teachers) the mean value of a lower level explanatory variable (an analytical variable). However, other aggregation functions may also be useful. For instance, one may have the hypothesis that classes that are heterogeneous with respect to some variable differ from more homogeneous classes. In this case, the aggregated explanatory variable would be the group's standard deviation or the range of the individual variable. Another aggregated value that can be useful is the group size (which is a global variable).

In SPSS/PC+, aggregation is handled by the procedure AGGREGATE. This procedure produces a new file that contains the grouping variable and the (new) aggregated variables. A simple setup to aggregate the variable IQ in a file with grouping variable GROUPNR is as follows:

```
GET FILE `indfile.sys'.
AGGREGATE outfile='aggfile.sys'/BREAK=groupnr/
  meaniq=MEAN(iq)/stdeviq=SD(iq).
```

Disaggregation means adding group level variables to the individual data file. This creates a file where the group level variables are repeated for all individuals in the same group. In SPSS/PC+, this can be accomplished by the procedure JOIN MATCH, using the so-called TABLE lookup. Before JOIN MATCH is used, the individual and the group file must both be sorted on the group identification variable. For instance, if we want to read the aggregated mean IQ and IQ

standard deviation to the individual file, we have the following setup:

```
JOIN MATCH FILE='indfile.sys'/
  TABLE='aggfile.sys'/BY groupnr/MAP.
```

The example below is a complete setup that uses aggregation and disaggregation to get group means and individual deviation scores for IQ:

```
GET FILE `indfile.sys'.
SORT groupnr.
SAVE FILE `indfile.sys'.
AGGREGATE outfile='aggfile.sys'/PRESORTED/BREAK=groupnr/
  meaniq=MEAN(iq)/stdeviq=SD(iq).
JOIN MATCH FILE='indfile.sys'/
  TABLE='aggfile.sys'/BY groupnr/MAP.
COMPUTE deviq=iq-meaniq.
save file `indfile2.sys'.
```

In this setup I use the AGGREGATE subcommand PRESORTED to indicate that the file is already sorted on the BREAK variable groupnr, because this saves computing time. The subcommand MAP on the JOIN MATCH procedure creates a map of the new system file, indicating from which of the two old system files the variables are taken. In this kind of `cutting and pasting' it is extremely important to check the output of both AGGREGATE and JOIN MATCH very carefully to make sure that the cases are indeed matched correctly.

It should be noted that the program HLM contains a built-in procedure for centering explanatory variables. The program MLn has a procedure to add group means to the individual data file, and commands to create centered and group-centered variables.

REFERENCES

Aiken, L.S., & West, S.G. (1991). *Multiple regression: Testing and interpreting interaction*. Newbury Park, CA: Sage.

Aitkin, M & N. Longford (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society*, 149, Part 1, pp. 1-43.

Aitkin, M., Anderson, D., Francis, B., & Hinde, J. (1989). *Statistical modelling in GLIM*. Oxford: Clarendon Press.

Alker, H.R. (1969). A typology of fallacies. In: M. Dogan & S. Rokkan (eds). *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, Mass.: M.I.T. Press.

Bangert-Drowns, R.L. (1986). Review of developments in meta-analytic method. *Psychological Bulletin*, 99, 388-399.

Bentler, P.M. *EQS. Structural equations program manual*. Los Angeles: BMDP Statistical Software, Inc.

Blalock, Hubert M., jr. (1979). Measurement and conceptualization problems: the major obstacle to integrating theory and research. *American Sociological Review*, 44: 881-894.

Blalock, Hubert M., jr. (1984). Contextual-effects models: theoretical and methodological Issues. *Annual Review of Sociology*, 10: 353-72.

Blalock, H.M. (1990). Auxiliary measurement theories revisited. In: J.J. Hox & J. De Jong-Gierveld (eds.). *Operationalization and Research Strategy*. Amsterdam: Swets & Zeitlinger.

Bock, D. (ed.) (1989). *Multilevel Analysis of Educational Data*. San Diego: Academic Press.

Boyd, Lawrence H., jr. & Iversen, Gudmund R. (1979. *Contextual Analysis: Concepts and Statistical Techniques*. Belmont, CA: Wadsworth Publ. Co.

Bryk A. S. & Raudenbush S.W. (1987). Applying the hierarchical linear model to measurement of change problems. *Psychological Bulletin*. 101, 147-158.

Bryk, A.S. & Raudenbush, S.W. (1988). Methodology for cross-level organizational research. In: *Research in Organizational Behavior*.

Bryk, A.S. & Raudenbush, S.W. (1992). *Hierarchical Linear Models*. Newbury Park, CA: Sage.

Bryk, A.S., Raudenbush, S.W., Seltzer, M. & Congdon, R.T. (1988). An introduction to HLM. Computer Program and User's Guide. Version 2.0.

Bryk, A.S., Raudenbush, S.W. & Congdon, R.T. (1994). HLM 2/3. *Hierarchical Linear Modeling with the HLM/2L and HLM/3L Programs*. Chicago: Scientific Software International.

Burstein, L. (1980). The analysis of multilevel data in educational research in evaluation. *Review of Research in Education*, 8, 158-233.

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

Burstein, L., Kim, K.-S., & Delandshere, G. (1989). Multilevel investigations of systematically varying slopes: Issues, alternatives, and consequences. In: Bock, D. (ed.) (1989). *Multilevel Analysis of Educational Data*. San Diego: Academic Press.

Burstein, L. (1978). Assessing differences between grouped and individual-level regression coefficients. *Sociological Methods & Research*, 7: 5-28.

Camstra, A. & Boomsma, A. (1992). Cross-validation in regression and covariance structure analysis: an overview. *Sociological Methods & Research*, 21, 89-115.
Cliff, N. (1987). *Analyzing Multivariate Data*. Orlando, Harbourt Brace Jovanovich.
Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.
Cohen, J. & Cohen, P. (1975, 1983). *Applied Multiple Regression Analysis for the Behavioral Sciences*. Hillsdale, NJ: Erlbaum.
Cook, T.D., & Weisberg, S. (1982). *Residuals and Influence in Regression*. New York & London: Chapman & Hall.
Cronbach, L.J. (1976). Research in classrooms and schools: formulation of questions, designs and analysis occasional paper: Stanford Evaluation Consortium.
Cronbach, L.J. & Webb, N. (1979). Between class and within class effects in a reported aptitude x treatment interaction: a reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology*, 67, 717-724.
De Jong-Gierveld, J. (1985). The development of a Rasch-type loneliness scale. *Applied Psychological Measurement*, 9, 289-299.
De Leeuw, E.D. (1992). *Data Quality in Mail, Telephone and Face to Face Surveys*. Amsterdam: TT-publikaties.
De Leeuw, E.D. & Van der Zouwen, J. (1988) Data quality in telephone and face-to-face surveys: a comparative meta-analysis. In: R. Groves et al. *Telephone Survey Methodology*. New York: Wiley.
De Leeuw, J. (1990). Data modeling and theory construction. In: J.J. Hox & J. De Jong-Gierveld (eds.) *Operationalization and Research Strategy*. Amsterdam: Swets & Zeitlinger.
De Leeuw, J. & Kreft, Ita G.G. (1986) Random coefficient models. *Journal of Educational Statistics*, 11, 1, 55-85.
DiPrete, T.A. & Grusky, D.B. (1990). The multilevel analysis of trends with repeated cross-sectional data. In: C.C. Clogg (ed.) *Sociological Methodology*, (1990. London: Blackwell.
Durkheim, Emile. (1898). *Suicide*. Translation published by The Free Press, Glencoe, Ill., 1951.
Erbring, L. & Young, A.A. (1979). Contextual effects as endogenous feedback. *Sociological Methods & Research*, 7:396-430.
Fotiu, R.P. (1989). A Comparison of the EM and Data Augmentation Algoritms on Simulated Small Sample Hierarchical Data from Research on Education. Unpublished doctoral dissertation, Michigan State University, East Lansing.
Galtung, J. (1969). *Theory and Methods of Social Research*. New York: Columbia University Press.
Goldstein, H. (1986). Efficient statistical modelling of longitudinal data. *Annals of Human Biology*, 13, 129-141.
Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. London: Griffin.
Goldstein, H. (1989). Efficient statistical modelling of longitudinal data. In Bock, R.D. (ed.): *Multilevel Analysis of Educational Data*. San Diego: Academic Press.
Goldstein, H. (1991) Non-linear multilevel models, with an application to discrete response data. *Biometrika*, 78, 45-51.
Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods & Research*, 22,

364-376.
Goldstein, H. (1995). *Multilevel Statistical Models*. London: Edward Arnold/New York: Halsted.
Goldstein, H. & McDonald, R. (1988). A general model for the analysis of multilevel data. *Psychometrika*, 53, 455-467.
Goldstein, H. & Silver, R. (1989). Multilevel and multivariate models in survey analysis. In: C. Skinner, D. Holt, & F. Smith (eds). *The Analysis of Complex Surveys*. New York: Wiley.
Goldstein, H., Healy, M.J.R. & Rasbash, J. (1994). Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, 13, 1643-1656.
Härnqvist, K., Gustaffson, J.E., Muthén, B.O. & Nelson, G. (1992). Hierarchical models of ability at individual and class levels. Submitted for publication.
Hays, W.L. (1973). *Statistics for the Social Sciences*. London: Holt, Rinehart & Winston.
Healy, M. (1987). Nanostat users guide (unpublished).
Hedeker, D. (1993). MIXOR. A Fortran Program for Mixed-effects Ordinal Probit and Logistic regression. Chicago: College of Medicine, University of Illinois.
Hedges, L.V. & Olkin, I. (1985). *Statistical Methods for Meta Analysis*. Orlando: Academic Press.
Hox, J.J. (1993). Factor analysis of multilevel data. Gauging the Muthén model. In: J.H.L. Oud & R.A.W. van Blokland-Vogelesang (eds.). *Advances in longitudinal and multivariate analysis in the behavioural sciences*. Nijmegen, NL: ITS.
Hox, J.J. (1994). Split2. Computer Program, Department of Education, University of Amsterdam.
Hox, J.J. & de Leeuw, E.D. (1986). De invloed van individuele en contextuele kenmerken op schoolbeleving en schoolprestatie. In: J.C. van der Wolf & J.J. Hox (eds). *Kwaliteit van Onderwijs in het Geding*. Amsterdam: Swets & Zeitlinger. (`The effect of individual and contextual variables on school attiude and learning'. In: The Quality of Education at Issue).
Hox, J.J., Kreft, Ita G.G. & Hermkens, P.L.J. (1991). The analysis of factorial surveys. *Sociological Methods & Research*, 19, 493-510.
Hox, J.J., de Leeuw, E.D., & Kreft, Ita G.G. (1991). The effect of interviewer and respondent characteristcs on the quality of survey data: a multilevel model. In: P. Biemer et al. (eds.). *Measurement Errors in Surveys*. New York: Wiley.
Hox, J.J. &. de Leeuw, E.D. (1994). A comparison of nonresponse in mail, telephone, and face to face surveys. *Quality & Quantity*, 28, 329-344.
Hummel, H.J. (1972. *Probleme der Mehrebeneanalyse*. Stuttgart: Teubner, (1972.
Hunter, J.E. & Schmidt, F.L. (1990). *Methods of Meta-analysis*. Newbury Park, CA: Sage.
Iverson, G.R. (1991). *Contextual Analysis*. Newbury Park, CA: Sage.
Jaccard, J., Turrisi, R. & Wan, C.K. (1990). *Interaction Effects in Multiple Regression*. Newbury Park, CA: Sage.
Kerlinger, F.N. & Pedhazur, E.J. (1973). *Multiple Regression in Behavioral Research*. New York: Holt, Rinehart & Winston.
Kirk, R.E. (1968). *Experimental Design: Procedures for the Behavioral Sciences*. Belmont, Calif: Brooks/Cole.
Kish, L. (1987). *Statistical Design for Research*. New York: Wiley.

Kreft, Ita G.G., & De Leeuw, E.D. (1987). The see-saw effect: a multilevel problem? A reanalysis of some findings of Hox and De Leeuw. *Qualtity & Quantity*, 22, 127-137.
Kreft, Ita G.G. & Kim, K.-S. (1990). A review of the statistical package ML3: Software for three level analysis by Rabash, Prosser and Goldstein. *Applied Statistics*, 40, 2, 343-346.
Kreft, Ita G.G., De Leeuw, J. & Kim, K.-S. (1990). Comparing Four Different Statistical Packages for Hierarchical Linear Regression: GENMOD, HLM, ML2, and VARCL. CSE Technical Report #310.
Kreft, Ita G.G. & de Leeuw, J. (1991). Model based ranking of schools. *International Journal of Educational Research*, 15, 1, 45-61.
Kreft, Ita G.G. & de Leeuw. J. (1993). The gender gap in earnings: A two-way nested multiple regression analysis with random effects. *Sociological Methods & Research*, 22, 319-341.
Lambert, P.C. & Abrams, K.R. (1995). Meta-analysis using multilevel models. *Multilevel Modelling Newsletter*, 7, 2, 17-19.
Lazarsfeld, P.F. & Menzel, H. (1961). On the relation between individual and collective properties. In: A. Etzioni (ed.). *Complex Organizations: A Sociological Reader*. New York: Holt, Rhinehart & Winston.
Lindley, D.V. & Novick, M.R. (1981). The role of exchangeability in inference. *Annals of Statistics*, 9, 45-58.
Longford, N.T. (1990) VARCL. Software for Variance Component Analysis of Data with Nested Random Effects (Maximum Likelihood), Educational Testing Service, Princeton, NJ
Longford, N.T. (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects. *Biometrika*, 74, 817-827.
Longford, N.T. (1988). A quasilikelihood adaption for variance component analysis. Educational Testing Service.
Longford, N.T. (1989a) Fisher scoring algorithm for variance component analysis of data with multilevel structure. in: Bock,D. R. (ed.) *Multilevel Analysis of Educational Data* . New York, Academic Press.
Longford, N.T. (1989b) To center or not to center in multilevel analysis. *Multilevel Modelling Newsletter*, vol.1 # 3 ,7,8 and 11
Longford, N.T. (1993). *Random Coefficient Models*. Oxford: Clarendon Press.
Longford, N.T. & Muthén, B. (1992). Factor analysis for clustered observations. *Psychometrika*, 57, 581-597.
Mason, W.M., Wong, G.M., & Entwisle, B. (1984). Contextual analysis through the multilevel linear model. In: S. Leinhardt (ed.). *Sociological Methodology, 1983-84*. San Francisco: Jossey-Bass.
McCullagh, P., & Nelder, J.A. (1883, (1989). *Generalized Linear Models*. London: Chapman & Hall.
McDonald, R.P. (1994). The bilevel reticular action model for path analysis with latent variables. *Sociological Methods & Research*, 22, 399-413.
McDonald, R.P. & Goldstein, H. (1989). Balanced versus unbalanced designs for linear structural relations in two-level data. *British Journal of Mathematical and Statistical Psychology*, 42, 215-232.

Mislevy, R.J. & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D. Bock (ed.) *Multilevel Analysis of Educational Data*. San Diego, CA: Academic Press.

Mosteller, F. & Tukey, J.W. (1977). *Data Analysis and Regression*. Reading, Mass: Addison-Wesley.

Muthén, B. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika*, 54, 557-585.

Muthén, B. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research*, 22, 376-398.

Muthén, B. & Satorra, A. (1989). Multilevel aspects of varying parameters in structural models. In: Bock, D. (ed.) (1989). *Multilevel Analysis of Educational Data*. San Diego: Academic Press.

Nelson, G. & Muthén, B. (1991). Analysis preparation steps for multilevel analysis using Liscomp. Technical Report, University of california, Los Angeles.

Nuttall, D.L., Goldstein, H., Prosser, R., & Rasbash, J. (1989). Differential school effectiveness. *International Journal of Educational Research*, 13, 764-776.

Pedhazur, E.J. (1977). Coding subjects in repeated measures designs. *Psychological Bulletin*, 84, 298-305.

Plewis, I. (1989) Comment on `Centering' Predictors in Multilevel Analysis. *Multilevel Modelling Newsletter*, vol.1 # 3 ,.6 and 11.

Prosser, R., Rasbash, J., & Goldstein, H. (1991). *ML3 Software for Three-level Analysis. Users Guide*. London: Institute of Education, University of London.

Rasbash, J. & Woodhouse, G. (1995). *MLn Command Reference*. London: Multilevel Models Project, University of London.

Raudenbush, S.W. (1989b) A Response to Longford and Plewis. *Multilevel Modelling Newsletter*, vol.1 # 3, 8-10

Raudenbush, S.W. (1989a) Centering predictors in multilevel analysis: choices and consequences. Multilevel Analysis *Multilevel Modelling Newsletter*, vol.1 # 2 , 10-12

Raudenbush, S.W. & Bryk, A.S. (1985). Empirical Bayes meta analysis. *Jounal of Educational Statistics*, 10, 75-98.

Raudenbush, S.W. & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, 59, 1-17.

Raudenbush, S.W. & Bryk, A.S. (1987). Examining correlates of diversity. *Journal of Educational Statistics*, 12, 241-269.

Raudenbush, S.W. (1987). Educational applications of hierarchical linear models. A review. *Journal of Educational Statistics*. 13, 85-116.

Raudenbush, S.W. & Bryk, A.S. (1988). Methodological advances in studying effects of schools and classrooms on student learning. *Review of Research on Education*.

Roberts, K. & Burstein, L. (1980). *New Directions for Methodology for Social and Behavioral Sciences. Vol. 6*. San Francisco: Jossey-Bass.

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

Skinner, C.J., Holt, D., & Smith, T.M.F. (eds.) (1989). *Analysis of Complex Surveys*. New York:

Wiley.
Snijders. T.A.B. & Bosker, R.J. (1993). Standard errors and sample sizes for multilevel research. *Journal of Educational Statistics*,
9Snijders. T.A.B. & Bosker, R.J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, 22, 342-363.
Snijders, T.A.B., Spreen, M. & Zwaagstra, R. 1994. Networks of cocaine users in an urban area: the use of multilevel modelling for analysing personal networks. *Journal of Quantitative Anthropology*, 5, 85-105.
Stinchcombe, A.L. (1968). *Constructing social theories*. New York: Harcourt.
Stram, D.O., Laird, N.M., & Ware, J.H. (1986). An algorithmic approach to the fitting of a general mixed ANOVA model appropriate in longitudinal settings. *Computer science and statistics: Proceedings of the seventeent symposium on the interface*. Amsterdam: North Holland.
Swanborn, P.G. (1981). *Methoden van Socialwetenschappelijk Onderzoek*. (Methods of social research). Amsterdam/Meppel: Boom.
Van den Eeden, P., & Saris, W. (1984). Empirisch onderzoek naar multilevel uitspraken. (empirical research into multilevel assertions) *Mens en Maatschappij*, 59, 165-178.
Van den Eeden, P. & Hüttner, H.J.M. (1982). Multilevel research. *Current Sociology*, vol. 30, 3: 1-181.
Van Duijn, M., Snijders, T.A.B. & Lazega, E. (1994). Random effects models for directed graphs. (in preparation).
Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54, 426-482.
Woodhouse, G. (1995) (Ed.). *A Guide to MLn for New Users*. London: Multilevel Models Project, University of London.
Wong, G.Y., & Mason, W.M. (1985) The hierarchical logistic regression model for multilevel analysis. Extensions of the hierarchical Normal Linear Model for Multilevel Analysis. *Journal of the American Statistical Association*, 80, 513-524.
Wong, G.Y. & Mason, W.M. (1989). Ethnicity, Comparative Analysis, and a Generalization of the Hierarchical Normal Linear Model for Multilevel Analysis. Population Studies Center Research Report # 89-138. The Population Studies Center of The University of Michigan.

AUTHOR INDEX

TOPIC INDEX