This is to certify that the

dissertation entitled

A MODEL FOR MULTILEVEL PATH ANALYSIS

presented by

Frank F. Jenkins

has been accepted towards fulfillment
of the requirements for

PhD degree in Education

Stephen W. Rawlenbusch
Major professor

Date 11-7-88

A MODEL FOR MULTILEVEL PATH ANALYSIS

By

Frank Ford Jenkins

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational
Psychology and Special Education

1988

ABSTRACT

A MODEL FOR MULTILEVEL PATH ANALYSIS

By

Frank Ford Jenkins

In the past couple of decades there have emerged two particularly innovative trends in quantitative research methodology in Education, path analysis and multilevel linear models. In path analysis, networks of interrelationships among variables are posited to represent the interconnected relationships found in real life processes. In multilevel linear models, analysis techniques have been devised to represent real life processes as they naturally occur in hierarchically nested contexts. Both approaches seek to represent complexity in a way that corresponds to the complexity of nature. There is developed in this thesis a method which combines these two trends into a multilevel path analysis. Such an approach combines the descriptive power of both path analysis and multilevel linear models, resulting in a single model which can define a complex network of processes for numerous groups simultaneously. The development of multilevel path models shows promise to increase the descriptive power and theory building ability of social science research.

The multilevel path modeling approach I have developed is a direct extension of the recent developments in empirical Bayes multilevel regression models. This is a methodology

by which to represent path analysis within numerous groups. First of all a path model is stipulated for each group. It is assumed that the path coefficients vary randomly from group to group. This variation is modeled by a between-group regression in which group-level variables predict the path coefficients. Contextual variables at a higher level of aggregation are introduced as predictors to explain why processes vary from group to group. When the errors of the within-group structural equations are assumed to be orthogonal, estimates for the first-stage and second-stage parameters are available via the EM algorithm.

Two hierarchical datasets are analyzed using this technique. The results indicate that novel insights into sociological processes can be gained by employing multilevel path models.

This dissertation is dedicated to my mother and Robin and to the world which I didn't help improve while I worked on it.

ACKNOWLEDGMENTS

I wish to acknowledge the members of my committee:

Bill Schmidt, for his preserverence and longevity;

Joe Byers for asking the tough questions;

Richard Houang who was a creative sounding board during the critical derivational phase of the work, and a helpful friend during the whole project;

Steve Raudenbush who inspired the entire program conceptually as well as personally. He encouraged my initial wild speculation and helped me tame my wild prose.

I would also like to acknowledge my friends and family who cajoled and endured. Finally I wish to remember the three jokers in the Friday morning group.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I:

STRATEGIES FOR MULTILEVEL DATA

I. Introduction

Educational researchers are faced with the task of studying quite complex phenomena. Students in classrooms possess a varied and unique personal histories. These histories interact with a vast array of inherited traits to form a matrix of propensities that the researcher must unravel. In addition to the complexity of the individual, researchers find that students function within a complicated hierarchy of social institutions: students are grouped into classes, classes are nested within schools, schools are nested within communities and so on to national and international levels.

Instead of bracketing out the complexity of educational contexts through tightly controlled 'laboratory' experiments, educational researchers have usually opted to capitalize on the rich variability found in schools. By studying natural settings researchers have hoped to address, and offer solutions to, problems as they naturally occur in schools.

In the past couple of decades there have emerged two particularly innovative trends in quantitative research methodology in education that deal with complexity, path analysis and multilevel linear models. In path analysis, networks of interrelationships among variables are posited to represent the interconnected relationships found in real

life processes. In multilevel linear models, analysis techniques have been devised to represent real life processes as they naturally occur in hierarchically nested contexts. Both approaches seek to represent complexity in a way that corresponds to the complexity of nature. As of yet, both approaches have represented disparate lines of inquiry informing and speaking to each other very little. In this thesis there will be developed a method which combines these two trends into a multilevel path analysis. Such an approach combines the descriptive power of both path analysis and multilevel linear models, resulting in a single model which can define a complex network of processes for numerous groups simultaneously. The development of multilevel path models shows promise to greatly increase the descriptive power and theory-building capacity of social science research.

Path analysis has traditionally emphasized the need for rich substantive theory. With origins in macro economics (Theil, 1971) and sociology (Duncan & Featherman, 1973) it has often been used to model large scale systems, e.g. the economy, ignoring smaller subunits. For example, a national analysis might not separately analyze each state economy. The focus in path analysis is usually on the need to extract a rich set of variables relevant to the processes being modeled. In education, sociology and economics, path models are applied as if one homogenous group were being studied. The reality of groups imbedded in a hierarchy of social structure is, for convenience, ignored. This oversight can

occur in two ways. In the first case what is regarded as a single uniform group might actually be composed of numerous dissimilar social units. For instance, students in a school might be regarded as a homogenous group for the purposes of a study while the fact is ignored that students are actually grouped into classroom units within the school. The opposite sort of oversight that can occur is the case in which the researcher ignores the fact that the group under study is one of numerous social units and each unit might exhibit different relationships among educational processes. For example, a single classroom might be studied ignoring the fact that effects estimated for that class might not generalize to other classes due to differences in the classroom context. In both of these cases if group membership of subjects were adequately taken into account, it would be possible to model processes within groups and then explore the generalizibility of effects across groups.

Educational researchers have long been concerned with multilevel issues, But traditional research methods have not provided adequate tools with which to analyzed data arising in naturally occurring hierarchies. The paradigm of educational research has been borrowed from the traditions of agriculture and psychology in which subjects are randomly assigned to each of several treatment conditions (Raudenbush & Bryk, 1988). This assures that the expected effect of confounding factors is zero. In addition the researcher, if possible, administers treatment to each subject individually,

thus assuring that the responses of one subject is independent of the responses of another subject (Raudenbush & Bryk, 1988).

Most educational research deviates from this paradigm in both of its aspects. Usually students are not randomly assigned to groups such as classrooms or schools and groups are usually not randomly assigned to treatments. Unfortunately, researchers find that they cannot control for confounding factors occurring both at the group and individual level, "The problem is that the statistical methods the educational researcher has inherited from experimental psychology provides little guidance on how to implement such statistical controls" (Raudenbush & Bryk, 1988).

The problem is further exacerbated by the fact that usually the independent factor of interest, or the 'treatment', is not administered individually to each student. For example, school factors effect all students within the school at the same time. In another example a classroom treatment often is administered to all the students in a class simultaneously. If students are affected as a group by independent factors, they will to some extent have a common group history and group experience. As a result, group responses will tend to be correlated, not independent. This lack of independence of responses violates statistical assumptions in traditional linear models.

Because of the problems of analyzing multilevel data by traditional methods a growing number of methodologists have

recognized the need to develop new research models which are multilevel in character.

The roots of multilevel linear models go back to Lindley and Smith (1972) and Smith (1973) and the definition of the General Bayesian Linear Model, which defines a linear model at multiple levels. Later researchers have capitalized on this theme in research in multilevel contexts (Rubin, 1980; Strenio, 1981; Morris, 1983; and others). A single model using an empirical Bayes approach for a wide range of educational applications was reviewed by Raudenbush (1984).

The multilevel path modeling approach developed in this thesis draws its inspiration from the notion of "slopes as outcomes" developed by Burstein and others (Burstein, Linn & Capell, 1978) and is a direct extension of the empirical Bayes estimation of a multilevel regression model reviewed by Raudenbush (1988). What is being proposed is a methodology by which to represent path analysis within numerous groups. First of all, a path model is stipulated for each group. It is assumed that the path coefficients vary randomly from group to group. In previous multilevel models it has been assumed that processes are homogenous across groups (Wisenbaker & Schmidt, 1979; Houang & Schmidt, 1981). The variation of processes is modeled by a between-group regression in which group-level variables predict the variance of the path coefficients. This between-group regression is a way to explicitly address heterogenous effects across groups. When variables and the paths among them are properly specified,

group-level processes which vary from group to group provide a source of explanation rather than constituting a violation of a homogeneity assumption.

II. PROBLEMS WITH MULTILEVEL CONTEXTS

As was pointed out above, path analysis research and traditional educational research have tended to ignore the nesting of individuals within groups. In nested contexts the lack of random assignment leads to confounded effects and the group-wide administration of treatment leads to correlated responses. As a result four major problems arise in the traditional analysis of hierarchically nested data.

1. In analysis of variance studies, the assumption of the independence of the errors of the units of analysis is violated making statistical tests invalid. This will result in an inflated actual alpha level, so that "The result ... is an unacceptably high type I error rate." (Barcikowski, 1981). This occurs because the precision of the effect is misestimated. Precision will be overestimated as a function of sample size and intraclass correlation (Walsh, 1947). Test statistics that are not corrected for this inflation will have high type I error rates whenever the intraclass correlation is greater than zero.

2. Aggregation bias can occur so that estimated relationships at one level of aggregation may be quite different from those at another level. This most commonly

occurs when the grouping variable is related to the outcome (Cooley, Bond & Mao, 1981). For example, consider a regression analysis predicting student achievement from student SES. Suppose data is aggregated to class means and average achievement is regressed on average SES. If students are tracked into classes according to pretest achievement, the regression coefficient estimated from class means will probably be much larger than that estimated from individual scores.

Aggregation bias may also occur simply because processes at one level are different from processes at another level (Burstein, 1980: Cooley, Bond & Mao, 1981). This could be because " Variables have different meanings at different levels of analysis." (Burstein, 1980) For example, a variable may gauge a student's desire to work alone. At the student level the variable may measure autonomy and motivation. But aggregated to the class level the variable may indicate group divisiveness.

Alternatively, there may be factors at one level of aggregation that are absent at another level and which moderate the process being studied. For example, a school district might provide low achieving classes with tutors to coach classes in the state achievement exam. Low achieving classrooms would increase in mean achievement, thus attenuating the SES/achievement relationship at the class mean level. But within each classroom, student SES might still have a large relationship with achievement.

3. Cross-level interactions alter individual level relationships from group to group. Cronbach and Webb (1975) realized that characteristics of the classroom interacted with and altered processes occurring among individuals. He saw this as a barrier to the formulation of scientific theories. Cronbach believed that the interaction of the setting with treatment made each study unique rendering it impossible to generalize beyond each setting and blocking the establishment of general theory.

4. Concentrating on one level of analysis loses information at the other level, and ignores cross-level interactions. Researchers can, for example, pool data within groups which enables them to take out the mean group effect and thus ignore group boundaries. This approach ignores possible setting-by-treatment interactions and assumes all groups have identical within-group processes. This of course is only tenable when effects really do happen to be uniform across groups. At the other extreme, researchers can aggregate to group means, which leads to the problem described by Page: "More rigorous investigators are apt to suppress most of the richness within the classroom by using class means" (1975 p.339). In this case possible cross-level interactions are ignored as are problems of aggregation bias. Single-level analyses must assume that effects are homogenous over groups. It would be better not make such a restrictive assumption and model the variation that occurs in effects from one group to another by group-level variables.

III. PATH ANALYSIS AND MULTILEVEL CONTEXTS

Path analysis offers promise for establishing theories in complex social contexts. A network of variables tested in a path analysis has an exact correspondence to a network of processes proposed by substantive theory. But path analysis has a history of largely neglecting the issue of hierarchical nesting of subjects. One of the earliest attempts to address this issue is found in Schmidt (1969) who articulated a maximum likelihood technique for partitioning a covariance matrix into orthogonal within-group and between-group parts. This approach was later extended by Wisenbaker and Schmidt (1979) to include structural models. A major limitation of these techniques is that they required that the group sizes be equal, constraining the practical applications.

Bianchi (1987) devised a Bayesian estimation technique which employed the EM algorithm (Dempster, Laird Rubin, 1977) to provide maximum likelihood estimates of latent random effects in the case of unequal group size. With the unequal-n solution in hand, it is possible to apply a partitioned covariance structure solution to natural settings.

Houang and Schmidt (1981) surveyed various methods of partitioning estimates into orthogonal between-group and within-group components in a context mostly pertinent to regression applications. They devised an analytical model which encompassed most of the partitioning approaches but they constrained within-group effects to be uniform across groups..

These approaches have remedied the first two problems hierarchical data described above; the proper estimation of precision and the avoidance of aggregation bias. However the problem of cross-level interactions, was not addressed by these models. Because all of these approaches partition structural parameters into independent within-group and between-group parts, the within-group partition is a single set of parameters using information pooled from each of the groups. This pooling of information is predicated upon the assumption that the within-group parameters are identical for all groups. Also, as they have been defined in the literature, these models do not allow for group level variables that are not also defined at the individual level.

IV. EDUCATIONAL RESEARCH AND MULTILEVEL CONTEXTS

Investigators using quasi-experimental designs have become increasingly aware of the problems for analysis posed by multilevel data. As early as 1940 McNemar recognized the problem of inflated alpha level for statistical tests, although he did not articulate the causes of what has come to be known as the "unit of analysis" problem. In educational research the problem has taken on the aspect of a Devil's bargain, as stated by Glass and Stanley (1970, p.507), "The researcher has two alternatives, though he is seldom aware of the second one: (1) he can run a potentially illegitimate analysis on the experiment by using the 'pupil' as the unit of statistical analysis, or (2) he can run a legitimate

analysis on the means of the classrooms, in which case he is almost certain to obtain statistically nonsignificant results." In my terms, the bargain is this: We can ignore groups and get problems one and three, cited above, or we can aggregate and get problems two, three and four.

More recently, researchers have begun to realize that they need not accept this no-win approach to research. Glass and Stanley framed the multilevel problem purely in terms of problem 1, improper ANOVA estimates. Hopkins (1982) sought to remedy the dilemma by devising a mixed model ANOVA approach for individuals nested within groups, and groups nested within treatments. Barcikowski (1981) explicitly defined the relationship between group size, intraclass correlation, actual effect size, and power in the ANOVA context. Cooley, Bond and Mao (1981) explain the origin of several species of aggregation bias and suggest multilevel structural equation modeling, of sorts, to remedy the situation.

Most educational researchers have not addressed the issue of cross-level interactions in multilevel contexts. Cronbach and Webb (1975) recognized the salience (and magnitude) of this issue. They contended that characteristics of the research setting (or, group-level variables) interact with within-group treatment effects, destroying the external validity of most quasi-experiments. Taking this lead, Burstein, Linn and Capell (1978) addressed the problem of multilevel data in terms of regression analysis. They suggested a model in which

regression parameters vary from group to group. In this way the variability of processes could be defined in the model, obviating the need to assume homogeneity across groups. Moreover, Burstein et al, proposed the notion of "slopes as outcomes", that is, using group characteristics as predictors for the within-group slopes. In the simplest case, the relationship between an outcome and a predictor is represented by a within-group regression weight, or slope. The slopes from all groups are then treated as outcomes for a second-stage analysis. In the second-stage analysis group-level variables predict the slopes in a multiple regression. Hanushek (1974) proposed using slopes as outcomes as a means of combining numerous regression studies, but his approach required slopes to be statistically independent. Although Burstein, Linn and Capell realized the implication of this approach for explicating cross-level interactions they did not delineate the statistical properties of the least squares estimates they proposed (Houang and Schmidt, 1981), leaving issues of statistical assessment of estimators and variance accounted for unresolved.

A problem with slopes-as-outcomes models which is even more serious than the lack of an overall statistical framework is the fact that slopes are very unreliable. The sampling variance of beta weights is usually much larger than sampling variance of ordinary outcomes, such as means. This unreliability will often mask the effect of group-level

predictors (Raudenbush and Bryk, 1986), washing out the information to be gleaned from modeling the slopes.

Another statistical problem has to do with the fact that the slopes vary in precision from group to group. For the second-stage analysis the slopes are outcomes to be analyzed by an ordinary least squares procedure that assumes equal precision for each slope. As a result of the violation of this assumption, the second-stage estimation procedure is less efficient leading to less precise second-stage parameter estimates. Because of this, it is more difficult to demonstrate the relationship between group-level variables and slopes (Raudenbush & Bryk, 1986).

A final problem has to do with the fact that the variability of slopes can be partitioned into two components; parameter variance, which represents the real differences in the slope parameters from group to group, and sampling variance, which is the error in slope estimates due to sampling. Only the parameter variance can be explained by a between-group model. For example, a between-group model which explains only a small portion of the total variance may in fact explain virtually all of the parameter variance. Unless parameter and sampling variance can be distinguished it will be very difficult to assess how well the between-group model accounts for slopes. The slopes-as-outcomes model does not provide means for partitioning slope variance (Raudenbush & Bryk, 1986).

V. THE HIERARCHICAL BAYESIAN LINEAR MODEL

A statistical theory which had promise to flesh out the slopes as outcomes approach was initiated when Lindley and Smith (1972) used Bayesian theory to provide alternatives to least squares estimates of the general linear model. The result was a hierarchical Bayesian linear model (Smith, 1973) in which structural parameters could be estimated for a two-stage hierarchy. These derivations assume that the dispersions and structural coefficients of the prior distributions are known.

In general, what the family of Bayesian linear models provide is a scheme in which a model can be specified in two stages. The first stage describes the data, given first stage parameter vector, B;

$$Y = XB + R,$$

with "X" containing the fixed predictors and "R" containing the random errors. The second stage describes the first stage slope parameters, given second stage parameters γ;

$$B = W\gamma + U,$$

with "W" being fixed predictors and "U" being random errors. A third stage defines the second stage parameters;

$$\gamma = AC + L.$$

This third stage simply defines our prior degree of certainty about the value of γ (Smith, 1973). The variance of the errors are;

$$\mathrm{Var}(R) = \Psi,$$

$$\mathrm{Var}(U) = T,$$

$$\mathrm{Var}(L) = \Gamma .$$

The goal is to get Bayesian or posterior estimates of the first and second level parameters (B and γ), given Y and C. Note that X, W, C, Ψ, T and Γ are assumed to be known.

What makes these models peculiarly Bayesian in character is the notion that the first and second stage parameters, B and γ, have distributions. When the data are normal these distributions can be described by linear models. The motivation for shifting focus from classical estimators is twofold. First, under most conditions the posterior estimates of parameters have smaller expected means squared error than classical estimators (Efron and Morris, 1977). Second, the notion of parameters as random latent variables can be conceptually appealing in multilevel contexts where within-group effects are commonly seen to differ from one group to another across a population of groups.

The hierarchical Bayes linear model proposed by Lindley and Smith fits rather naturally into a notion like "slopes as outcomes" where the first stage parameter vector, B, is interpreted as the within-group slopes, and the second-stage parameter vector, γ, is interpreted as the between-group regression coefficients of "B" predicted by group-level variables found in "W".

An example of an application of the hierarchical Bayesian linear model is found in Rubin (1981). This is an instance

of an empirical Bayesian application of this model to educational research. Empirical Bayes is different from pure Bayes in that prior distributions of parameters are estimated from the data, instead of being given. Although Rubin assumed prior dispersions of the data to be known, he used the data to estimate the prior location and dispersion of the second-stage parameters. This study demonstrated how Bayes and empirical Bayes techniques can give superior estimates of treatment effects by combining information from within-group and between-group sources.

Rubin (1981) had used a graphical method to get a maximum likelihood estimate for second-stage dispersions, which was an option available in the simplified model he employed. Generally, though, there was at the time no uniform method for estimating prior dispersions. In order to apply the hierarchical Bayesian linear model one had to work out a solution pertinent only to a specific, less complicated case.

VI. EMPIRICAL BAYES THROUGH THE EM ALGORITHM

Widespread acceptance of the EM algorithm led to a practical approach for estimating prior dispersions. Dempster, Laird and Rubin (1977) outlined a general formulation of the EM algorithm, an iterative computational method which yields maximum likelihood estimates for a wide variety of estimation problems. It was termed "EM" because each iteration consists of an Expectation phase followed by a Maximization phase. The power of this algorithm is that it will give estimates for

'incomplete' data when one specifies maximum likelihood equations for 'complete' data. In linear models with normally distributed data the 'incomplete' data consist of the observed outcomes. The 'complete' data include the observed data plus certain latent variables, e.g. second-stage errors. If the complete data were observed it would be a simple matter to obtain maximum likelihood estimates for dispersions. By acting 'as if' one had complete data one can greatly simplify maximization equations. In the expectation phase dispersions are treated as known quantities, and the "complete data sufficient statistics" needed for the M step are estimated.

Generally, the algorithm works like this: In the "E" step, parameter estimates from the previous iteration are used to calculate the conditional expected value of the "complete data" sufficient statistics, given the observed (incomplete) data. So in this step, sufficient statistics are derived as if parameters were known. In the "M" step, the sufficient statistics from the "E" step are used to calculate the maximum likelihood estimates of parameters. So estimates of parameters are derived as if the complete data had been observed.

By bouncing back and forth between the "E" and the "M" step, the likelihood converges to a maximum. If the algorithm is applied to data that is normally distributed, it should converge to a global maximum and yield asymptotically efficient ML estimates (Raudenbush, 1984).

VII. THE HIERARCHICAL LINEAR MODEL

The general formulation of the EM algorithm paved the way for the widespread implementation of the empirical Bayesian linear model. Strenio (1981) first implemented the EM algorithm for this purpose. Raudenbush (1984) devised a mixed model empirical Bayes approach he called the Hierarchical Linear Model, or HLM. This is a very flexible model that can be tailored to apply to three, heretofore disparate, realms of research; School effects research with regression modeling, meta-analysis with treatment effects, and growth curve estimation for individual students. Other investigators have used the EM algorithm in mixed linear models (see Laird & Ware, 1982; Strenio, Weisberg & Bryk, 1983; and Mason, Wong, & Entwisle, 1984). Other approaches to dispersion estimation have been proposed by Goldstein (1986), and Longford (1985). Raudenbush (1988) reviews these developments.

In hierarchical linear models for school effects research, there is a single regression model which is estimated in numerous groups. This regression model is usually characterized by there being one predictor of interest and several covariates which control for the confounding effects of student background characteristics.

In hierarchical linear models for meta-analysis there is a set of separate studies all focusing on a similar 'treatment' issue. The effect from each group is usually a standardized mean difference between treatment and non-treatment groups.

The HLM of individual growth studies (see Bryk & Raudenbush, 1987) look at student growth over time on an outcome of interest, controlling for background characteristics of the individuals.

Raudenbush (1988) demonstrates how the slopes as outcomes interpretation of the hierarchical Bayesian linear model can elucidate all of these research contexts. This unified approach to multilevel analysis is characterized by a) heterogenous effects b) separate estimates of sampling variance and parameter variance of the first level parameters c) posterior estimates which offer smaller expected mean squared error than corresponding least squares estimates and d) between-group predictor coefficients of the first level parameters. This model speaks to all four problems of multilevel analysis that have been outlined.

VIII. HIERARCHICAL PATH MODELS

There is an important limitation with the HLM approach: it only uses one type of model (multiple regression) to depict within-group processes. In the research paradigms where the primary interest is with the relationship between several independent variables and one dependent variable, multiple regression is conceptually appealing. A multiple regression depicts a 'many-to-one' type of relationship, as shown in figure 1.1.

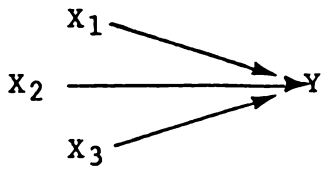


Figure 1.1

A Many-To-One Relationship

If this relationship represents the actual processes according to substantive theory, a multiple regression defines a structural equation. But if there are multiple interrelated outcomes, multiple regression (and ANCOVA, which can be put in terms of multiple regression) defines a prediction relationship only and causal imputations can be extremely misleading. For example, suppose that in actual classrooms, enjoyment of reading contributes to reading comprehension, and comprehension contributes to reading achievement:

Enjoyment ——————> Comprehension ——————> Achievement

A regression analysis predicts achievement by enjoyment and comprehension to give the following:

$$\text{Achievement} = B_0 + B_1\,(\text{Enjoyment}) + B_2\,(\text{Comprehension}).$$

To impute a direct effect of enjoyment on achievement from what could be a large regression coefficient would quite distort the picture. Researchers often wish they could draw

structural conclusions from predictive equations, and not a few succumb to the temptation to do so (if only in the sanctuary of their own thoughts).

Muthen and Satorra (1987) surveyed numerous modeling issues connected with multilevel structural models. They broached the questions of 1) heterogenous group parameters and 2) correlated within-group responses. The discussion ranges over a wide variety of issues to do with assumptions of the nature of regressors, (fixed, random or latent) and whether various parameters are homogenous or heterogenous across groups. When they considered strategies for estimating the models they defined, they were less than optimistic:

> "It is clear that today's standard structural equation modeling techniques and software cannot fully serve the purposes of an appropriate multilevel analysis." (p.19)

In this thesis the challenge will be taken to start filling in the gap between theory and implementation for multilevel structural models.

The model being proposed here deals with a particular subclass of models considered by Muthen and Satorra. In the first stage a path model is defined within each group. The path model is the same for all groups but the path coefficients can randomly vary from group to group. A single group model for group j is given by;

$$Y_j = Z_j B_j + R_j,$$

where R represents a vector of random errors for group j, Z_j is a vector of fixed predictors which includes the endogenous predictors found in Y_j, and B_j is a vector of random path coefficients.

The first-stage parameters are modeled at the second stage by a between-group regression model of the form:

$$B = W\gamma + U,$$

Where B is the vector of path coefficients from all groups, W is a set of fixed between-group predictors, γ is a set of fixed between-group coefficients and U is a set of random errors.

In terms of Muthen and Satorra (1987) several features define what subclass of possible hierarchical structural models this is:

1) The within-group predictors, Z_j, are fixed.

2) The within-group coefficients, B_j, are random and heterogenous across groups.

3) The variance of the second-stage errors, $Var(U_j) = \tau$, is homogenous across groups.

4) The second stage predictors, W, are fixed.

5) The second-stage parameters, γ, are fixed.

Other choices were possible for each of the five options indicating that there are a large number of different hierarchical models that can be devised.

An hidden feature of this model is the structure that defines the path analysis, ZB. This appears to be identical to a regression model. But the matrices are specially

constructed so that "ZB" represents a series of structural equations stacked on top of each other. Details of this structure will be discussed in chapter two.

IX. DEMONSTRATING THE MODEL

In subsequent chapters the mathematical model for the hierarchical path analysis will be defined. Then the computer algorithm that was developed to implement the model will be discussed. Finally, the efficacy of the model for explicating educational research will be demonstrated by using the model to analyze two actual data sets.

The efficacy of the model for explicating educational research will be demonstrated by presenting the analysis that has been done on two actual data sets. The first analysis was drawn from a large scale research project called the High School and Beyond study (Coleman, Hoffer & Kilgore, 1982). This study measured variables at the student level and at the school level in nearly a thousand schools across the country. The students were measured on mathematics achievement and various background variables. It has been found by various researchers that the relationship between student background (SES and race) and achievement is less strong in Catholic high schools than in public high schools (Coleman, Hoffer & Kilgore, 1982: Hoffer, Greeley & Coleman, 1985). A conclusion that has been drawn is that Catholic schools are more egalitarian than public schools since academic success depends less on students' background in Catholic schools.

Data from a random sample of 158 schools will be analyzed. The purpose of the analysis is to explain why Catholic schools appear more egalitarian than public schools, or rather to identify the school level characteristics account for the discrepancy between public and private schools with respect to the relationship of students' background to students' achievement.

In a second demonstration of the multilevel path model variables from a study by Willms (1987) will be reanalyzed. This data consists of observations taken from 21 secondary schools in one administrative division in Scotland. Measures were taken on various student background variables. Students' academic success was gauged by verbal reasoning score and a score on a comprehensive achievement exam. School means obtained by aggregating the data the group level were introduced to measure school context. The original study by Willms examined mean student achievement, controlled for by student background and verbal reasoning.

In the present analysis the student level processes will be construed as a network of causal relationships in which student background affects academic achievement indirectly through students verbal ability. The within-school path coefficients will be modeled by a between-school regression in which school context predicts variations in the processes by which students achieve educational goals.

It is hoped that these two analyses will demonstrate that hierarchical path models are feasible and that they can

significantly add to our understanding of educational processes in their full multilevel contexts.

CHAPTER II:

ESTIMATING MULTILEVEL PATH MODELS

I. Introduction

In this chapter estimators for parameters of the general Bayesian linear model will be derived. The general Bayesian model takes the individual form;

$$Y = A\theta + R, \qquad (2.1)$$

where;

Y - is a K by 1 vector of outcomes for an individual, with K = the number of outcomes;

A - is a K by s matrix of predictors, where s is the number of structural parameters;

θ - is an S by 1 vector of parameters;

R - is a K by 1 vector of random errors.

This model is Bayesian because the parameters in θ are assumed to be random terms with a prior probability distribution.

It will be assumed θ has a normal prior distribution with mean zero and dispersion matrix Ω. In Bayes terms this represents our prior belief about the location of the parameter vector, θ, and the precision of this prior belief (represented by Ω^{-1}). Estimation in the Bayesian context involves calculating the posterior distribution of the parameters, that is finding the posterior density function $f(\theta|Y)$. After the formula for the posterior distribution of θ is derived, the mean

vector of this posterior distribution will be used as the vector of point estimates of the vector parameter.

By an application of Bayes theorem to continuous probability densities it can be shown that the posterior density function is proportional to the product of two independent density functions; $f(\theta|Y)$ and $f(\theta)$ (and a constant term which drops out). This gives rise to the proportional relationship, signified by $\propto$ (see Hoel, Port and Stone, 1971, section 6.3);

$$f(\theta|Y) \propto f(Y|\theta)f(\theta) \ . \qquad (2.2)$$

The first term, $f(Y|\theta)$ represents the likelihood of the data, and the second term, $f(\theta)$, represents the prior distribution of the parameters. This division is fortuitous because it enables one to develop the likelihood and the prior distribution separately and bring the results together in one expression. This greatly simplifies the exposition.

An expression for posterior density of θ given Y can be had rather straightforwardly by substituting the normal probability density functions for $f(Y|\theta)$ and $f(\theta)$. Then by multiplying, combining and simplifying terms a probability density function, $f(\theta|Y)$, results which is recognizably normal. This expression will reproduce the standard Bayesian results for the General Bayesian Linear Model. This general solution is not hierarchical, i.e. it doesn't define parameters at two levels of analysis.

In order to define a hierarchical linear model we must reparameterize, using the substitutions;

$$A = [ZW : Z] , \tag{2.3}$$

and,

$$\Theta = \begin{vmatrix} \gamma \\ --- \\ U \end{vmatrix} \tag{2.4}$$

By substituting A and Θ into Equation 2.1 we get;

$$Y = ZW\gamma + ZU + R . \tag{2.5}$$

By introducing an identity for a new parameter, B, we can decompose the model into two stages. The identity is,

$$B = W\gamma + U .$$

Substituting this into Equation 2.5 leads to a two-stage expression,

$$Y = ZB + R \tag{2.6}$$

and,

$$B = W\gamma + U . \tag{2.7}$$

A convenient interpretation for this two-equation expression is that forms a two-stage hierarchy in which $Y = ZB + R$ represents a linear model within-groups and $B = W\gamma + U$ represents a between-groups linear model for the within-group parameter vector, B (following Smith, 1973).

In this thesis the within-groups model represents a path model defined within numerous groups in which set of paths for

each group j, B_j, differs from group to group. The between-group model is a multiple regression in which group-level variables, W, predict the group paths, B_j. The hierarchical Bayes model enables us to model processes at the within-group and between-group levels of analysis simultaneously, which is the strength of the multilevel modeling approach.

In this chapter the Bayesian estimates of parameters will first be derived for the general Bayesian linear model of Equation 2.1. It will be shown that the general Bayesian results are valid for the substituted or 'mixed model' case represented by Equation 2.5. Finally by switching to the two-stage model in Equations 2.6 and 2.7, it will be shown that by stipulating a recursive path model at the within-group level, we can justify the assumptions made in deriving Bayesian estimates.

Throughout the derivation of the estimates it is assumed that first and second stage variance matrices are known. This is usually an untenable assumption. For this reason, the EM algorithm, an empirical estimating routine, is used to provide maximum likelihood estimates of the variance terms. In the last section of this chapter I derive the likelihood of the data, conditioned on the variance parameters. It is this likelihood which is maximized by the EM algorithm. The derivation of the formulas used in the EM algorithm will be developed in chapter 3.

II. The General Bayesian Model

The general Bayesian linear model (Smith, 1973) as depicted in Equation 2.1 was defined for one individual. We will now define the model in terms of a whole group of N individuals with K outcomes per individual,

$$Y = A\Theta + R,$$

where,

Y is a KNx1 vector of outcomes,

A is a KNxQ matrix of predictors,

Θ is a Qx1 vector of random structural coefficients,

R is a KNx1 vector of random errors,

N is the total number of individuals,

K is the number of outcomes observed for each individual, and

Q is the number of parameters in the model.

The variance of the errors is,

$$Var(R) = \Psi \qquad (2.8)$$

where,

Ψ is a NK by NK variance matrix.

The sampling errors, R, are independent of the parameters represented by Θ. As a result Θ is assumed to have a normal prior distribution,

$$\Theta \sim N(0, \Omega) . \qquad (2.9)$$

where,

0 is a zero vector representing the prior mean and

Ω is the QxQ prior dispersion matrix.

The assumption of a zero prior mean can be made without loss of generalizability (Raudenbush, 1984).

In order to find Bayesian point estimates of a model, one must first derive an expression of the posterior distribution of the parameters given the data and conditional on the prior distribution of the parameters. In terms of the general Bayesian linear model, if we have a prior normal distribution of parameters, the posterior distribution has the form;

$$(\theta \mid Y,\Omega) \sim N(\theta^*, D_{\theta}^*), \qquad (2.10)$$

where θ^* is the posterior mean and $D_{\theta*}$ is the posterior variance matrix (Raudenbush, 1984). An explanation of how to find a posterior distribution by employing Bayes theorem, the heart of Bayesian estimation theory, is given in the following section.

Recall from Equation 2.2 that the posterior density is proportional to the product of the likelihood of the data and the prior density of θ or,

$$f(\theta|Y) \propto f(Y|\theta)\, f(\theta) \quad .$$

First we will focus on the likelihood, $f(Y|\theta)$. Conditional on θ, the errors of the observations will be independent across individuals. Also, if we assume that we have a well-specified, recursive structural equation system, the errors for K outcomes observed for each individual will also be independent (Land, 1973). This latter assumption

can be explicitly justified when we couch the model in hierarchical terms in a later section. Under these assumptions Land (1973) has shown that the joint normal probability density can be depicted as the product of the densities of the NK separate observations,

$$f(y_{i1},\ldots,y_{iK},\ \ldots\ ,y_{N1},\ldots,y_{NK}) = \prod_{i=1}^{N} \prod_{k=1}^{K} f(y_{ik}) \ . \qquad (2.11)$$

When the errors between individuals and between measures are uncorrelated, the variance of the errors, Ψ, is an NK by NK diagonal matrix. When this is the case, the probability density function depicted in Equation 2.11 takes the specific form,

$$f(Y|\theta,\Psi) =$$
$$(2\pi)^{-NK/2}\ |\Psi|^{-1/2}\ \exp\{-1/2(Y - A\theta)'\ \Psi^{-1}\ (Y - A\theta)\}. \qquad (2.12)$$

This completes half the task, that of defining the likelihood of Y. The prior distribution of θ follows from the assumptions that θ is normally distributed with zero mean and dispersion matrix Ω,

$$f(\theta) = (2\pi)^{-Q/2}\ |\Omega|^{-1/2}\ \exp\{-1/2(\theta'\ \Omega^{-1}\ \theta)\}\ , \qquad (2.13)$$

All that remains now is to combine $f(Y|\theta)$ with $f(\theta)$ to get $f(\theta|Y)$. We accomplish this by multiplying Equation 2.12 by 2.13 to get,

$$f(\theta|Y,\Psi,\Omega,A) \propto$$
$$(2\pi)^{-KN/2}\ (2\pi)^{-Q/2}\ |\psi|^{-N/2}\ |\Omega|^{-1/2}$$
$$\exp\{-1/2(Y-A\theta)'\ \Psi^{-1}\ (Y-A\theta)\}\ \exp\{-1/2\ \theta'\ \Omega^{-1}\ \theta\}\ . \quad (2.14)$$

By expanding, combining terms, completing the square in the quadratic term and combining constant terms (see appendix), Equation 2.14 can be shown to be proportional to the following density:

$$f(\theta|Y,\Psi,\Omega,A) \propto$$
$$\exp\{-1/2[\theta-(A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y]'\ (A'\Psi^{-1}A + \Omega^{-1})$$
$$[\theta-(A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y]\} \quad (2.15)$$

This is a normal density function with the mean and covariance of θ clearly visible. The posterior distribution of θ is therefore defined as follows;

$$(\theta|Y,\Psi,\Omega,A) \sim N(\theta^*, D_\theta^*)\ , \quad (2.16)$$

with, $\theta^* = (A'\Psi^{-1}A + \Omega^{-1})\ A'\Psi^{-1}Y$, and

$D_\theta^* = (A'\Psi^{-1}A + \Omega^{-1})$.

This reproduces standard Bayesian results (see Raudenbush, 1988). The estimate of θ will simply be the mean of the posterior distribution of θ or, $(A'\Psi^{-1}A + \Omega^{-1})A'\Psi^{-1}Y$, from Equation 2.16.

III. The Bayesian Mixed Model

We can write the mixed model form of the Bayesian model by substituting Equation 2.7 into Equation 2.6 to yield;

$$Y = ZW\gamma + ZU + R \ .$$

The first level parameter matrix, B, has disappeared so that the parameters of this model are γ and U. The parameters have the prior normal distributions,

$\gamma \sim N(0 \ , \ \Gamma)$, and

$U \sim N(0 \ , \ T)$.

The prior distribution for γ serves the purpose of representing our state of knowledge about γ (Smith, 1973). We assume γ has an arbitrarily large variance matrix so that the precision of γ, or Γ^{-1}, is near zero. This indicates a complete lack of knowledge about the prior distribution of γ. Such a distribution has appropriately been termed a vague prior (Smith, 1973). The mean of γ is set to zero for convenience, since with a vague prior the location of the parameter is arbitrary (Raudenbush, 1984).

Because there is no prior information about γ, it is functionally equivalent to a fixed effect , while U is considered to be a random effect (Dempster, Rubin & Tsutakawa, 1981). The combined model, then, loses the hierarchical character and can be thought of as a one level mixed model. Both of these conceptualizations, hierarchical and mixed model, will be used in this chapter. It is important to

keep in mind that both of these conceptualizations are equivalent.

The parameters of the mixed model, γ and U, are assumed to be independent so that the joint prior covariance of the parameters equals,

$$\text{Var} \begin{vmatrix} \theta_1 \\ \hline \theta_2 \end{vmatrix} = \begin{vmatrix} \Gamma & 0 \\ 0 & T \end{vmatrix} = \Omega. \qquad (2.17)$$

Mixed Model Posterior Estimates

The estimating algorithm used in this thesis was designed in terms of the Bayesian mixed model. This is because the mixed model affords two statistical advantages over a hierarchical model; it allows for groups to be analyzed which have data matrices that are not full rank and it allows for a flexible definition of which effects are fixed and which effects are random. These issues will be taken up in the concluding chapter.

A more general way to represent this model is to posit the following definitions,

$A_1 = ZW$; $A_2 = Z$; $\theta_1 = \gamma$; and $\theta_2 = U$.

Substituting these into Equation 2.5 we get the general form for the mixed model,

$$Y = A_1\theta_1 + A_2\theta_2 + R, \qquad (2.18)$$

where θ_1 is the vector of fixed effects and θ_2 is the vector of random effects.

A simple substitution show that there is a correspondence between the mixed model and the General Bayesian model. If we set,

$$A = [A_1 : A_2] \text{ ,and} \tag{2.19}$$

$$\theta = \begin{vmatrix} \theta_1 \\ \hline \theta_2 \end{vmatrix} , \tag{2.20}$$

and substitute into Equation 2.18, we see that the mixed model is just a partitioning of the General Bayesian model,

$$Y = A\theta + R \ .$$

Also note that the variance of θ, Ω, is simply the joint covariance of θ_1 and θ_2 as defined in Equation 2.18,

$$\text{Var } (\theta) = \text{Var} \begin{vmatrix} \theta_1 \\ \hline \theta_2 \end{vmatrix} = \begin{vmatrix} \Gamma & 0 \\ 0 & T \end{vmatrix} .$$

The formula for the posterior mean in Equation 2.16, calls for the inverse of Ω which is,

$$\begin{vmatrix} \Gamma^{-1} & 0 \\ 0 & T^{-1} \end{vmatrix} .$$

Recall that the prior precision of γ is Γ^{-1}, which is virtually zero. This leads to the results,

$$\Omega^{-1} = \begin{vmatrix} 0 & 0 \\ 0 & T^{-1} \end{vmatrix} . \tag{2.21}$$

With these definitions in hand, we can now substitute Equations 2.19, 2.20 and 2.21 into Equation 2.16 to get an expression of the posterior distribution of parameters in terms of the general mixed model.

By using various identities and simplifications it is possible to derive the following mixed model definition of the posterior distribution of parameters (see Raudenbush, 1988).

$$\left(\begin{vmatrix} \theta_1 \\ \hline \theta_2 \end{vmatrix} \;\middle|\; (Y,\ \psi,\ \Omega,\ A_1,\ A_2) \right) \sim N \left(\begin{vmatrix} \theta_1^* \\ \hline \theta_2^* \end{vmatrix} ;\ \begin{vmatrix} D_{11} & D_{12} \\ \hline D_{21} & D_{22} \end{vmatrix} \right) \tag{2.22}$$

With:

$$\theta_1^* = D_{11}A_1'(I - \Psi^{-1}A_2C^{-1}A_2)\,\Psi^{-1}Y$$

$$\theta_2^* = C^{-1}A_2'\Psi^{-1}\,(Y - A_1\theta_1^*)$$

$$C = A_2'\Psi^{-1}A_2 + T^{-1}$$

$$D_{11} = (A_1'\Psi^{-1}A_1 - A_1\Psi^{-1}A_2C^{-1}A_2'\Psi^{-1}A_1)$$

$$D_{12} = D_{21}' = - D_{11}A_1'\Psi^{-1}A_2C^{-1}$$

$$D_{22} = C^{-1} + C^{-1}A_2'\Psi^{-1}A_1D_{11}A_1'\Psi^{-1}A_2C^{-1} .$$

The mixed model solution subsumes the special case in which the model is strictly hierarchical as in Equation 2.10;

$$Y = ZW\gamma + ZU + R ,$$

where Z is a matrix of first-stage predictors and W is a matrix of second-stage predictors. This is equivalent to the general mixed model in Equation 2.18 only if $ZW = A_1$, $Z = A_2$, $\gamma = \theta_1$, and $U = \theta_2$. When these conditions are met ZW, Z, γ, and U can be substituted into Equation 2.16 to yield simpler equations for θ_1^*, θ_2^*, and D.

Simpler equations are desirable but there are important cases in which A_2 does not equal Z. For this reason we will adhere to the more general mixed model approach.

IV. Likelihood for the Hierarchical Case

In the derivation of posterior estimates for the general Bayesian linear model, it was assumed that the K outcomes observed for each individual had independent errors. This assumption enabled us to devise a simple expression for the likelihood of the data given parameters (Equation 2.12). In order to justify this assumption we must appeal to the hierarchical form of the Bayesian model.

In the hierarchical path model a vector of first-stage paths, B, is introduced as the set of parameters,

$$Y = ZB + R,$$

$$B = W\gamma + U .$$

The probability distribution function of the data conditional on B is $f(Y|B)$, which is the likelihood of the data parallel to $f(Y|\theta)$. Since the only random terms that both θ and B contain are γ and U, conditioning on B is equivalent to conditioning on θ. For this reason the derivation in this

section of an expression for the likelihood $f(Y|B)$ will be pertinent to the expression for the general Bayesian likelihood in Equation 2.12.

Conditional on first-stage paths, B, the model Y = ZB + R represents a path model with fixed exogenous predictors in Z and fixed paths, B. This defines an ordinary, non-Bayesian path model.

Structure of the First-Stage Path System

In order elucidate the structure of the First-stage path model we will examine a simple example in which the model Y = ZB + R represents a two equation path system for an individual, as illustrated by the path diagram in figure 2.1,

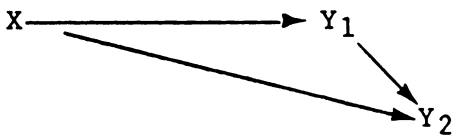


Figure 2.1
Path Diagram for a Two Equation System

In this example X is an exogenous variable because it has no antecedents in the path system, while Y_1 and Y_2 are endogenous variables because they both have causal antecedents defined in the system.

The structure of the matrix equation which depicts this path system is illustrated by figure 2.2.

$$\begin{matrix} Y & = & Z & & B & + & R \end{matrix}$$

$$\begin{vmatrix} Y_1 \\ Y_2 \end{vmatrix} = \begin{vmatrix} X & 0 & 0 \\ 0 & Y_1 & X \end{vmatrix} \begin{vmatrix} B_{x1} \\ B_{y1} \\ B_{x2} \end{vmatrix} + \begin{vmatrix} R_1 \\ R_2 \end{vmatrix}$$

Figure 2.2
Matrix Structure of a Two Equation Path System

This figure shows that each row of Z contains the predictors for one equation. Note that exogenous path are subscripted with x and endogenous path are subscripted with y. If we perform the matrix post-multiplication of Z with B and properly add elements we see that the matrix equation, Y = ZB + R is a shorthand way of writing the two equations,

Equation 1 $\quad Y_1 = XB_{x1} + R_1$

Equation 2 $\quad Y_2 = Y_1B_{y1} + XB_{x2} + R_2$.

Recursive Path Models

Throughout this thesis focus will be restricted to the subclass of path models which are recursive. A recursive path model is one in which there are no causal loops. A simple recursive path model is illustrated by the path diagram in figure 2.3.

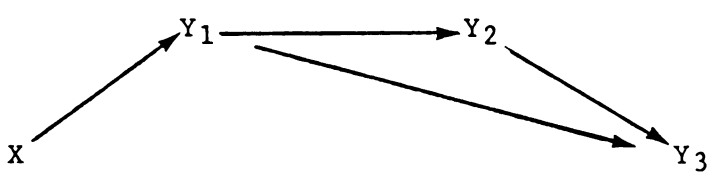


Figure 2.3
Example of A Recursive Path System

The arrows stand for causal connections between the processes represented by X, Y_1, Y_2, and Y_3. In this case X is an exogenous variable because no antecedents are defined for it. The endogenous variables are the Y's because they have causal antecedents defined within the model. Notice that the causal flow is in one direction, from left to right.

A non-recursive path system is illustrated by figure 2.4. We arrive at this model by simply reversing the direction of the arrow between Y_1 and Y_3.

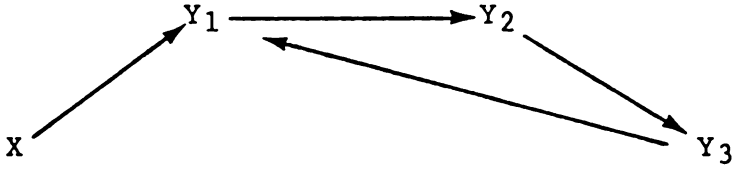


Figure 2.4
Example of A Non-Recursive Path System

If you trace the causal flow from Y_1 to Y_2 to Y_3, you will find that you will loop back to Y_1 again. Any such loop makes a path model non-recursive.

The reason we are limiting ourselves to <u>recursive</u> path models is that they define a class of models that have readily interpretable parameters which are also easily estimated. Fortunately, recursive path models also describe most processes of interest in the social sciences. Hunter and Gerbing (1980) states that in any non-recursive path system can be transformed into a recursive one if it is represented

longitudinally, because causal paths cannot loop backward in time.

Recursive path models have a certain general structure which will be exploited in the derivation below. To see the general structure we must define a full recursive model. If we add some paths to figure 2.1 we will get the full path model in figure 2.5.

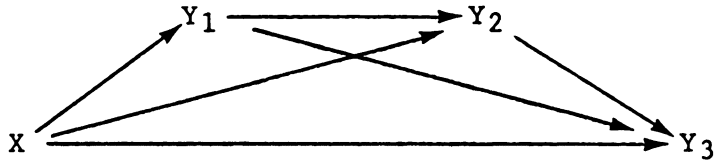


Figure 2.5
Example of A Full Recursive Path System

This model is full in the sense that if any path is added, the model becomes non-recursive. The individual level equation system represented in figure 2.6 illustrates the structure.

$$Y_1 = XB_{x1} + R_1$$

$$Y_2 = Y_1B_{y1} + XB_{x2} + R_2$$

$$Y_3 = Y_1B_{y2} + Y_2B_{y3} + XB_{x3} + R_3$$

Figure 2.6
Individual Level Equation System - Full Recursive Path Model

All of the terms are scalars. The paths for the endogenous and exogenous variables have been differentiated. The B_x's are the exogenous path coefficients and the B_y's are the endogenous path coefficients. Notice that the equations compound in a stepwise fashion. Each equation has as

predictors a) the exogenous variable and b) all of the endogenous variables in the equations above it. Every recursive system can be ordered in such a way as to display this structure. If we take the endogenous predictors and put them in a matrix as they appear in the equation system, we get the structure in figure 2.7.

$$B_y = \begin{vmatrix} 0 & (\text{zero}) \\ B_{y1} & 0 \\ B_{y2} & B_{y3} \quad 0 \end{vmatrix}$$

Figure 2.7
Structure of Endogenous Paths in a Full Recursive System

From this we see that in recursive path system the matrix of endogenous predictors will be a lower triangular matrix. This is another way to define a recursive system. This fact will come in handy when defining the likelihood function of Y in the next section. Note that a less-full recursive system would have some of the paths missing, so some of the B_y's in figure 2.7 would be set to zero. The general structure would be preserved, though.

Change of Variable for the Probability Density Function of Y

A strong assumption of the individual path model is that the errors of the equations, R_k, are uncorrelated. This assumptions simplifies the probability density function

(PDF) of Y given B because the dispersion matrix of errors for the individual, ψ, will be diagonal. In order to exploit this fact, the PDF of Y must be expressed in terms of R, which in calculus terms means there must be a change of variables for the PDF. Standard calculus theory tells us that with vector variables, in order to express a function of one variable in terms of a function of another variable, one must multiply by the determinant of the Jacobian of the transformation (Hoel, Port & Stone, 1971). In terms of the probability density functions in question we have,

$$f(Y|B) = f(R|B) \quad |\delta R / \delta Y| \qquad (2.23)$$

where, $f(Y|B)$ is the density in terms of the error vector R and $\delta R/\delta Y$ is the Jacobian of the transformation.

In order to express the density in terms of R we will have to restructure the model somewhat. First we will put the matrix equation in terms of R,

$$R = Y - ZB \ . \qquad (2.24)$$

As figure 2.1 shows each row of Z contains the endogenous and exogenous variables for one equation. For this reason, some Y_k's can appear in both the Y vector and the Z matrix. In order to differentiate (Y - ZB) with respect to Y (which the Jacobian requires) we need to have all of the Y_k's in one vector. This can be accomplished by transforming the right hand side of Equation 2.7 into a restructured but equivalent form. To see how this could be done we will rearrange the scalar terms from figure 2.6 and add some zero terms,

$$Y = ZB + R \qquad \text{Matrix Form}$$

$$Y_1 = 0Y_1 + 0Y_2 + 0Y_3 + B_{x1}X + R_1$$

$$Y_2 = B_{y1}Y_1 + 0Y_2 + 0Y_3 + B_{x2}X + R_1 \qquad \text{Single Equation Form}$$

$$Y_1 = B_{y2}Y_1 + B_{y3}Y_2 + 0Y_3 + B_{x3}X + R_1$$

Figure 2.8
Augmented Single Equation Form of Path Model

This suggests an alternative but equivalent structure in which the endogenous paths are grouped together in one matrix and the exogenous paths are grouped together in another matrix;

$$Y = B_y Y + B_x X + R$$

$$\begin{vmatrix} Y_1 \\ Y_2 \\ Y_2 \end{vmatrix} = \begin{vmatrix} 0 & 0 & 0 \\ B_{y1} & 0 & 0 \\ B_{y2} & B_{y3} & 0 \end{vmatrix} \begin{vmatrix} Y_1 \\ Y_2 \\ Y_3 \end{vmatrix} + \begin{vmatrix} B_{x1} \\ B_{x2} \\ B_{x3} \end{vmatrix} X + \begin{vmatrix} R_1 \\ R_2 \\ R_3 \end{vmatrix}$$

Figure 2.9
Restructured Single Equation Form of Path Model

Notice that the endogenous path matrix, B_y, displays the lower diagonal structure indicative of a recursive path system.

If we multiply this matrix equation and combine the proper terms we will get the same three equation system in figure 2.6. This demonstrates that the two alternative structures are equivalent,

$$ZB + R = B_yY + B_xX + R \ .$$

This restructured form provides a convenient expression for the model in terms of R,

$$R = Y - B_y Y - B_x X \quad .$$

Combining terms gives us,

$$R = (I - B_y)Y - B_x X \quad . \qquad (2.25)$$

This expression for R is also equivalent to the original expression in Equation 2.23,

$$Y - ZB = (I - B_y)Y - B_x X \quad .$$

Now Equations 2.24 and 2.25 can be substituted into Equation 2.23 to yield an expression of the PDF in terms of R,

$$f(Y|B) = f(Y - ZB|B) \; |\delta[(I - B_y)Y - B_x X]/\delta Y| \quad . \qquad (2.26)$$

The simplification afforded by a recursive is apparent if we focus on the Jacobian. Differentiating gives the result,

$$\delta[(I - B_y)Y - B_x X]/\delta Y = I - B_y \quad .$$

The results, $I - B_y$, is a lower triangular matrix which reflects the structure of recursive path systems,

$$I - B_y = \begin{vmatrix} 1 & 0 & 0 \\ -B_{y1} & 1 & 0 \\ -B_{y2} & -B_{y3} & 1 \end{vmatrix} \quad .$$

Because of this structure, the determinant of the matrix is unity (Land, 1973). Substituting this result into Equation 2.26 yields,

$$f(Y|B) = f(Y-ZB|B) * 1$$
$$= f(Y-ZB|B) \qquad (2.27)$$

The normal probability density for one individual follows straightforwardly from this form. It is assumed that 1) there are K outcomes for an individual and 2) the error vector for an individual is distributed, $R \sim N(0,\psi)$, where ψ is a K by K diagonal matrix since the errors of the equations are uncorrelated, 3) this is a well-specified recursive path system which explains why the errors are uncorrelated (Heise, 1975). The density according to Equation 2.27 and these assumptions is,

$$f(Y_i|B) = (2\pi)^{-K/2} |\psi|^{-1/2} \exp\{-1/2(Y-B)' \psi^{-1} (Y-B)\}. \qquad (2.28)$$

The Whole-Group Likelihood

Equation 2.28 gives us the density function for one individual. Assuming that each individual's response is independently and identically distributed, the whole-group likelihood will simply be the product of each individual's likelihood. If there are N individuals in the group this leads to the total-N PDF

$$f(Y|G,H,\psi) =$$
$$(2\pi)^{-NK/2} |\psi|^{-N/2} \exp\{-1/2\textstyle\sum_{(i)} (Y-B)' \psi^{-1} (Y-B)\}. \qquad (2.29)$$

where, i is the index for individuals and,

$\sum_{(i)}$ is the summation over all individuals.

An important feature of this model is that, since the dispersions are independent across equations, each equation of the multiple equation system can be estimated by an independent regression analysis. Also, the paths estimated in such a model are full information maximum likelihood estimates, as established by Land (1973). This will be important when it comes to implementing estimates with the EM algorithm, as we will see in chapter three.

The Bayesian Likelihood for Many Groups

The distinguishing feature of the hierarchical Bayesian model is that the structural coefficients differ from group to group in a way that is described by the second-stage model. First let us introduce a structural coefficient vector, B_j, that differs over group, j. The whole group coefficient vector, B, now represents the parameters from all groups stacked up vertically,

$$B = \begin{vmatrix} B_1 \\ B_2 \\ . \\ . \\ . \\ B_J \end{vmatrix} .$$

Conditional on B, individuals are independent within and between groups. As a result the total-N likelihood is the product of the individual likelihood functions within and between groups. This results in the addition of a second

group summation to Equation 2.29. If $N = \sum n_j$, is the total N for all groups, the many group likelihood becomes,

$$f(Y|B,\psi) = (2\pi)^{-NK/2} |\psi|^{-N/2} \exp\{-1/2 \sum_{(j)} \sum_{(i)} (Y_{ij} - Z_{ij}B_j)' \psi^{-1} (Y_{ij} - Z_{ij}B)\}. \quad (2.30)$$

where $\sum_{(i)}$ is the summation over individuals in a group and $\sum_{(j)}$ is a summation over groups.

This formula can be rendered in a simpler form without the summations. Since individuals are independent (conditioned on B) the whole group error variance matrix, Ψ, is a KN by KN block diagonal with identical blocks equal to ψ. Since the equations are uncorrelated each ψ is diagonal. So Ψ is also a diagonal matrix. For this reason the double summation in the exponent of Equation 2.30 can be rewritten in terms of whole group matrices,

$$f(Y|B,\Psi) = (2\pi)^{-NK/2} |\Psi|^{-1/2} \exp\{-1/2(Y - ZB)' \Psi^{-1} (Y - ZB)\}. \quad (2.31)$$

where Y is KN by 1, Z is KN by JP and B is JP by 1. P is the number of paths in the within-group model.

Transforming the Model Into the General Bayesian Likelihood

We have described the likelihood for Y conditional on first-stage parameter matrix B. The likelihood we ultimately seek is defined in terms of Θ. Recall that Θ is a partitioned matrix containing γ and U. We can introduce these parameters into the likelihood by recalling the identity for B found in the second-stage of the hierarchical model, $B = W\gamma + U$.

By substituting this identity into Equation 2.31, we get the substituted form of the likelihood in terms of γ and U,

$$f(Y|\gamma,U,\Psi) =$$
$$(2\pi)^{-NK/2} |\Psi|^{-1/2} \exp\{-1/2(Y-ZW\gamma-ZU)' \Psi^{-1} (Y-ZW\gamma-ZU)\}.$$

Now use the identities;

$$A = [ZW : Z], \text{ and } \quad \Theta = \begin{vmatrix} \gamma \\ --- \\ U \end{vmatrix} .$$

Substituting into the last equation yields the General Bayesian form,

$$f(Y|\Theta,\Psi) =$$
$$(2\pi)^{-NK/2} |\Psi|^{-1/2} \exp\{-1/2(Y - A\Theta)' \Psi^{-1} (Y - A\Theta)\}. \quad (2.32)$$

This is the likelihood that was used to derive the posterior estimates for the General Bayesian model. By showing that the hierarchical likelihood leads to the general Bayesian likelihood, we have proven that by assuming a recursive first-stage path model one can derive the general Bayesian estimates (Equation 2.16).

V. The Matrix Structure of the Hierarchical Bayesian Model

The structure of the matrices has been illustrated only for the single individual model. When considering all individuals in many groups a special structure has been devised for the matrices by Strenio (1981) to make elements conformable in the hierarchical Bayesian linear model. The whole group model can be constructed by stacking the individual level matrices in an appropriate manner. For example, if there are K outcomes and J groups and n_j individuals in a group, the whole group data vector, Y, is a $K\sum n_j$ by 1 vector which results from vertically stacking the $\sum n_j K$ by 1 individual outcome vectors. The sum, $\sum n_j$, sums the number of all individuals over all J groups and will be referred to as N. The whole group data vector has the form;

$$Y = \begin{bmatrix} Y_{11} \\ Y_{21} \\ \vdots \\ Y_{n1} \\ \vdots \\ Y_{1J} \\ Y_{2J} \\ \vdots \\ Y_{nJ} \end{bmatrix} .$$

The error vectors are stacked in a similar way to yield a total-N error vector, R, with dimension N by 1;

$$R = \begin{bmatrix} R_{11} \\ R_{21} \\ \vdots \\ R_{n1} \\ \vdots \\ R_{1J} \\ R_{2J} \\ \vdots \\ R_{nJ} \end{bmatrix} .$$

The individual predictor matrices, Z_{ij}, are stacked vertically within each group and then each group's data matrix is arranged in a block diagonal;

$$Z_j = \begin{bmatrix} z_{1j} \\ z_{2j} \\ \cdot \\ \cdot \\ \cdot \\ z_{nj} \end{bmatrix} ,$$

with,

$$Z = \begin{bmatrix} z_1 & & & & \\ & z_2 & & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & & z_J \end{bmatrix} .$$

Finally, the P by 1 coefficient vectors for each group, B_j, can be stacked vertically to yield the JP by 1 whole group vector, B;

$$B = \begin{bmatrix} B_1 \\ B_2 \\ \cdot \\ \cdot \\ \cdot \\ B_J \end{bmatrix} .$$

Combining these matrices gives us the first-stage whole group model:

$$\underset{KN \times 1}{Y} = \underset{KN \times JP}{Z} \quad \underset{JP \times 1}{B} + \underset{KN \times 1}{R} . \qquad (2.33)$$

The second-stage model is stacked by a similar process. The P by 1 U_j error vectors are stacked just at the B_j vectors are;

$$U = \begin{bmatrix} U_1 \\ U_2 \\ \cdot \\ \cdot \\ \cdot \\ U_J \end{bmatrix} .$$

Similarly, the p by s W_j group-level data matrices are stacked vertically for each group to yield a Jp by s matrix W:

$$W = \begin{bmatrix} W_1 \\ W_2 \\ \cdot \\ \cdot \\ \cdot \\ W_J \end{bmatrix} .$$

But each group-level data matrix has a block diagonal structure, with each block corresponding to each first-stage coefficient, B_{pj};

$$W_j = \begin{bmatrix} w_1 & & & \\ & w_2 & & (zero) \\ & & \cdot & \\ & & & \cdot \\ & (zero) & & \cdot \\ & & & w_p \end{bmatrix} .$$

with $w_p = [w_1, w_2, \ldots, w_s]$. The diagonal block terms, w_p, are thus row vectors which can be of variable length. In this way a different set of group-level predictors can predict each path.

The structure of the γ vector is no different in the

whole group case. Combining these terms leads to the whole group second-stage model;

$$\underset{JP \times 1}{B} = \underset{JP \times S}{W}\ \underset{S \times 1}{\gamma} + \underset{JP \times 1}{U} \qquad (2.34)$$

The distributional assumptions are similar to the single group case;

$$R \sim N(0, \Psi) ,$$

$$U \sim N(0, T) ,$$

$$\gamma \sim N(0, \Gamma) .$$

In addition it is assumed that R, U and γ are mutually independent.

As before, Γ^{-1} is assumed to be arbitrarily close to a zero matrix. The structure of Ψ and T differs from the individual model. Since, conditional on B, individuals are independent within and between groups, and since it assumed that individual errors are identically distributed, Ψ will be a block diagonal matrix with each block equal to the individual K by K variance matrix, ψ;

$$\Psi = \begin{vmatrix} \underset{kxk}{\psi} & & (zero) \\ & \underset{kxk}{\psi} & \\ & & \ddots \\ (zero) & & \underset{kxk}{\psi} \end{vmatrix} .$$

$$Nk \times Nk$$

Due to the fact that errors are independent across equations, each ψ is a diagonal matrix of the form,

$$\psi = \begin{bmatrix} \sigma_1 & & (\text{zero}) \\ & \sigma_2 & \\ & & \ddots \\ (\text{zero}) & & \sigma_K \end{bmatrix} .$$

K x K

The structural coefficient vector for a group, B_j, are assumed to be independent and identically distributed across groups. As a result the whole group variance matrix T is a block diagonal with each block consisting of the identical variance matrix for a groups structural parameters, B_j;

$$T = \begin{bmatrix} \underset{px1}{\tau} & & (\text{zero}) \\ & \underset{px1}{\tau} & \\ & & \ddots \\ (\text{zero}) & & \underset{px1}{\tau} \end{bmatrix} .$$

JP x 1

This completes the exposition of the hierarchical Bayesian model.

VI. Likelihood of the Data

In the derivation for the mixed model posterior distribution, it is assumed that variance matrices, ψ and τ, are known. By employing the EM algorithm it is possible to get maximum likelihood estimates for ψ and τ, but it is necessary to have a criterion to judge whether the EM algorithm has converged to the ML estimates. Therefore the relevant likelihood of the data, $f(Y|\psi,\tau)$, must be monitored after each iteration of the algorithm (Dempster, Rubin and Tsutakawa, 1981). In this section I give a derivation of this likelihood.

An expression for the probability density function of Y given ψ and τ can be obtained from the densities already defined.

Let us consider several densities, all of which are conditioned on ψ and τ. Bayes theorem for probability densities gives us an expression for the posterior density of θ given Y, ψ and τ (Hoel, Port & Stone, 1971),

$$f(\theta|Y,\Psi,T) = \frac{f(\theta,Y|\psi,\tau)}{f(Y|\psi,\tau)} .$$

Solving for the PDF of Y given ψ and τ,

$$f(Y|\psi,\tau) = \frac{f(\theta,Y|\psi,\tau)}{f(\theta|Y,\psi,\tau)} . \qquad (2.35)$$

The likelihood we seek is $f(Y|\psi,\tau)$. We can cast the joint PDF, $f(\theta,Y|\psi,\tau)$, into a more useful form by noting that,

$$f(Y|\theta,\psi,\tau) = \frac{f(\theta,Y|\psi,\tau)}{f(\theta|\psi,\tau)} \quad . \tag{2.36}$$

Solving for $f(\theta,Y|\psi,\tau)$ yields an alternative form of the joint PDF,

$$f(\theta,Y|\psi,\tau) = f(Y|\theta,\psi,\tau)\ f(\theta|\psi,\tau) \quad . \tag{2.37}$$

Substituting Equation 2.35 into Equation 2.33 gives,

$$f(Y|\psi,\tau) = \frac{f(Y|\theta,\psi,\tau)\ f(\theta|\psi,\tau)}{f(\theta|Y,\psi,\tau)} \quad . \tag{2.38}$$

This is the same as the form given by Demptser, Rubin and Tsutakawa (1981).

The densities on the right hand side of Equation 2.38 have been defined in terms of normal density functions earlier in this chapter in Equations 2.12, 2.13 and 2.15 respectively. Substituting in to Equation 2.38 and eliminating some constant terms reveals the following likelihood,

$$f(Y|\Psi,T) \propto$$

$$\frac{|\Psi|^{-1/2}\exp\{-1/2(Y-A\theta)'\Psi^{-1}(Y-A\theta)\}|\Omega|^{-1/2}\exp\{-1/2(\theta'\Omega^{-1}\theta)\}}{|D_\theta^*|^{-1/2}\ \exp\{-1/2\ (\theta-\theta^*)'\ D_\theta^{*-1}\ (\theta-\theta^*)\}}$$

with, $D_\theta^* = (A'\Psi^{-1}A+\Omega^{-1})$ and $\theta^* = (A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y$ from Equation 2.41. Combining terms leads to,

$$|\Psi|^{-1/2}|\Omega|^{-1/2}|D_\theta^*|^{1/2}\exp\{-1/2(Y-A\theta)'\Psi^{-1}(Y-A\theta)+\theta'\Omega^{-1}\theta$$
$$- (\theta-\theta^*)D_\theta^{*-1}(\theta-\theta^*)\} \quad .$$

This expression holds for all θ, therefore a convenient simplification can be had by evaluating the expression at $\theta=\theta^*$ (see Dempster, Rubin & Tsutakawa, 1981). The exponential term reduces to,

$$\exp\{-1/2(Y-A\theta)'\Psi^{-1}(Y-A\theta)+\theta^*\Omega^{-1}\theta^*\} .$$

Using the fact that $\theta^* = D^*_\theta A'\Psi^{-1}Y$ this can be simplified to;

$$\exp\{-1/2Y\Psi^{-1}(Y-A\theta)\} .$$

Now the log can taken to yield the final form,

$$f(Y|\psi,\tau) \propto$$
$$-\text{Log}|\Psi| -\text{Log}|\Omega| \text{ Log}|D^*_\theta| -1/2\ Y'\Psi^{-1}(Y-A\theta^*) . \qquad (2.39)$$

This can be put in terms of the mixed model by substituting the following terms into Equation 2.39;

$$A = [A_1 : A_2],\ \theta^* = \begin{vmatrix} \theta^*_1 \\ \hline \theta^*_2 \end{vmatrix} ,\quad D^*_\theta = \begin{vmatrix} D_{11} & D_{12} \\ \hline D_{21} & D_{22} \end{vmatrix} .$$

As a result of these substitutions three terms in Equation 2.16 will change;

$$1)\quad -\text{Log}|\Omega| = -\begin{vmatrix} \Gamma & 0 \\ 0 & T \end{vmatrix} = -\text{Log}\ |\Gamma| -\text{Log}\ |T| .$$

Because γ is assumed to have a vague prior, its precision, Γ^{-1}, goes to zero. This implies that Γ is arbitrarily large and fixed (Dempster, Rubin & Tsutakawa, 1981), so $|\Gamma|$

is treated as a constant and is taken out of the effective part of the likelihood expression.

2) The term $\text{Log}|D^{*}_{\theta}|$ can be reexpressed using a standard method for taking the determinant of a partitioned matrix,

$$|D^{*}_{\theta}| = |D_{11}|\ |D_{22}-D_{21}D_{11}^{-1}D_{12}| \ .$$

By substituting the equivalencies for D_{22} and D_{21} from Equation 2.22 and simplifying, The expression reduces to,

$$|D^{*}_{\theta}| = |D_{11}|\ |C^{-1}|,$$

where $C^{-1} = (A_2'\Psi^{-1}A_2+T^{-1})^{-1}$ (Raudenbush, 1987-B).

3) Finally, $Y-A\Theta$ has the mixed model form $Y - A_1\theta_1 - A_2\theta_2$.

Substituting these changes into Equation 2.39 yields the mixed model log likelihood:

$$\text{Log } P(Y|\Psi,T) \propto$$
$$-\text{Log}|\Psi| - \text{Log}|T| + \text{Log}|D_{11}| + \text{Log}|C^{-1}| - Y\Psi^{-1}(Y-A_1\theta^{*}_1-A_2\theta^{*}_2). \qquad (2.40)$$

This last expression is used as the criterion for the EM algorithm which will be discussed in chapter three.

CHAPTER III

METHODS

I. Introduction

This chapter will review some of the technical aspects of implementing multilevel path analysis. First I review the EM algorithm and explain its rationale. Then the specific equations will be derived for implementing the EM algorithm in a multilevel context for estimating variance components. Next, statistical tests will be discussed. These will encompass the chi-square test of parameter variance, the R^2 statistic for assessing fit of the second-stage model and the Z test of second level parameters. Finally there will be a discussion about the validation of the computer algorithm. This will focus on the cross-referencing analysis that was done with the multilevel path analysis program and on the Hierarchical Linear Model program (Bryk, Raudenbush, Seltzer & Congdon, 1986).

II. Implementation of the EM Algorithm

The logic of the EM algorithm is to estimate parameters for a hypothetically complete set of data from a sample space which only contains incomplete data. Instead of using the actual summary statistics found in the data, which are by definition 'incomplete', the EM algorithm utilizes the expected value of complete data summary statistics as a

substitute for having the 'complete data' statistics. The advantage of this strategy is that maximum likelihood estimators based on the assumption of complete data can be quite simple to derive.

The EM algorithm is an iterative routine which cycles through an expectation phase and a separate maximization phase at each iteration. The maximization phase consists of the calculation of maximum likelihood estimates for parameters based on the assumption of complete data. Consider a simple example of variance estimation. Let us assume a model for individual i;

$$Y_i = X_i B + e_i \quad ,$$

where,

Y_i is a single outcome,

X_i is a matrix of fixed predictors,

B is a vector of regression coefficients and

e_i is the random error.

We want to estimate σ given that $Var(e) = \sigma^2 I$, the variance of the errors. If the complete data consists of observations of Y_i as well as of e_i, the maximum likelihood estimator of σ^2 is simply defined as,

$$\hat{\sigma}^2 = \sum e_i^2/N \quad ,$$

where, $\sum e_i^2$ is the "complete data sufficient statistic", that is, the sufficient statistic needed to obtain the ML estimate given that one has observed the complete data.

In actuality, though, we never observe the e_i's. With the EM algorithm we use the conditional expectation of the complete data sufficient statistic instead of the actual complete data sufficient statistic;

$$E(\sum e_i^2 \mid Y) ,$$

where Y is the incomplete, i.e. observed, data. The expected value of the sufficient statistics is calculated during the Expectation phase of the EM algorithm.

A general schema for the EM algorithm is,

$$P_1 = F\{E_\theta(\text{Sufficient Statistics}|P_0, \text{ Incomplete Data})\}$$

with,

P_0 = Vector of parameter estimates from the previous iteration of the algorithm,

P_1 = Vector of parameter estimates for the present iteration,

$F(\ )$ = The estimator of parameter vector P_1 assuming complete data.

E_P = Expectation over all possible values of P, given the complete data

Sufficient Statistics$|P_0$, Incomplete Data = sufficient statistics given previous estimates of parameters and the observed, incomplete, data.

Note that sufficient statistics are conditioned on the data. This means that parameter estimates involved in calculating sufficient statistics will be the Bayesian estimators derived in chapter two.

I will use the same model explicated in chapter two for the EM derivations which has the hierarchical form,

$$Y = ZB + R$$

$$B = W\gamma + U \ .$$

The substituted model is,

$$Y = ZW\gamma + ZU + R.$$

As before, by making the substitution of,

A_1=ZW; A_2=Z; θ_1=γ; and θ_2=U,

we get the mixed model form, $Y = A_1\theta_1 + A_2\theta_2 + R$.

Also as before, $Var(U_j)$=τ, and $Var(R_{ij})$=ψ, where U_j is the p by 1 vector of parameter errors for the paths of one group and R_{ij} is the k by 1 vector of sampling errors for person i in group j. The purpose of the EM algorithm is to estimate τ and ψ.

EM Formula for Estimating The Variance Matrix τ

In the context of estimating τ the 'complete data' consists of the observed outcome data, Y, and the second-stage errors, U_j. Assuming complete data the ML estimator for τ is simply,

$$\sum U_j U_j'/J \quad \text{(Raudenbush, 1987-B)},$$

with, U_j = The p x 1 vector of parameter errors associated with the p paths in group j.

J = The number of groups.

With the EM approach we substitute $E(U_jU_j'|Y,\tau_0,\psi_0)$ for U_jU_j', at each iteration (Dempster, Rubin, Tsutakawa, 1981).

The expected sufficient statistics for a vector product like U_jU_j', comes out of the definition of variance in standard statistics theory. The dispersion of U_j, $Var(U_j)$, is defined as,

$$Var(U_j) = E(U_jU_j') - E(U_j)E(U_j)' \text{ (Searle, 1971).}$$

Now we solve for the quantity we seek, the expected value of the sufficient statistic, $E(U_jU_j')$;

$$E(U_jU_j') = E(U_j)E(U_j)' + Var(U_j) \quad .$$

But the EM algorithm requires the expected sufficient statistics given the 'incomplete' data, Y and parameter estimates from the previous iteration, τ_0 and ψ_0, so the expectation is;

$$E(U_jU_j|Y,\tau_0,\psi_0) = U_j^*U_j^{*\prime} + D^*_{Uj} \ ,$$

where, U_j^* is the posterior parameter estimate of θ_2 given in chapter two in Equation 2.22, and

D^*_{Uj} is the posterior dispersion matrix for θ_2, also given in Equation 2.22.

The connection between the posterior estimates given known variances developed in chapter two, and the EM estimating routine is now explicit. At each iteration you plug in the

variance estimates from the previous iteration as the 'known' variance and then you use the formulas for posterior estimates developed in chapter two to calculate the expected sufficient statistics.

The maximization phase for the estimation of τ is accomplished by the trivial operation of dividing the expected sufficient statistics by j,

$$\tau_1 = \sum(U_j^* U_j^{*\prime} + D_* U_j)/J \ .$$

This completes one iteration for estimating τ.

EM Formula for Estimating First-Stage Variance Matrix ψ

The first-stage variance term, ψ, is a K by K diagonal matrix with off diagonals of zero. Because the first-stage errors are uncorrelated, the variance terms can be estimated by K separate EM estimation calculations. The K separate estimates are then arranged along the diagonal of ψ to provide the matrix estimate. Each of the K parallel estimates follows the same format.

The quantity to be estimated in one EM calculation is the k,k scaler diagonal element of ψ, σ^2_k. The 'complete data' in this case consist of the N by 1 observed outcome vector, Y_k, and the N by 1 first stage error vector, R_k. The complete data maximum likelihood estimate for σ^2_k is,

$$R_k' R_k = \sum_{(j)} \sum_{(i)} R^2_{ijk}/N \ ,$$

where, R_k is the N by 1 error vector for variable k,

R_{ijk} is the error for person i, group j, and outcome (endogenous variable), k,

$N = \sum_{(j)} n_j$, is the total number of individuals in all groups, and

The double summation indicates summing the squared errors over persons and groups.

The derivation of the expected sufficient statistics in the case of σ^2_k is more complicated than for τ. The term R_k is the N by 1 error vector for outcome k. The mixed model formula with R_k is,

$$Y_k = A_{1k}\theta_{1k} + A_{2k}\theta_{2k} + R_k \ .$$

We solve for R_k to get,

$$R_k = Y_k - A_{1k}\theta_{1k} - A_{2k}\theta_{2k}$$

So the complete data sufficient statistics for σ^2_k is,

$$\sum_j \sum_i R^2_{ijk} =$$

$$(Y_k - A_{1k}\theta_{1k} - A_{2k}\theta_{2k})'(Y_k - A_{1k}\theta_{1k} - A_{2k}\theta_{2k}) \ .$$

This last formula can be made more tractable by putting the mixed model into its simpler General Model form using the substitution,

$$A = [A_1 : A_2], \text{ and } \theta_k = \begin{vmatrix} \theta_{1k} \\ \hline \theta_{2k} \end{vmatrix} \ ,$$

the sufficient statistic now has the form,

$$\sum_j \sum_i R^2_{ijk} =$$

$$(Y_k - A_k\theta_k)'(Y_k - A_k\theta_k) \ .$$

The expected value for a scalar quadratic of the form $R_k'R_k$ is,

$$E(R_kR_k') = E(R_k)'E(R_k) + Tr\{Var(R_k)\},$$

were, $Tr\{Var(R_k)\}$ is the trace of the dispersion matrix for the N by 1 vector, R_k.

But for the EM algorithm we need the expectation given ψ_0, τ_0 and Y. In the first term we have,

$$E(R_k|\psi_0,\tau_0,Y_k) = Y_k - A_k\theta^*_k,$$

where, θ^*_k is the estimate of the posterior parameter mean for equation k, found in Equation 2.22 of the last chapter.

The second term, $Var(R_k)$, is the posterior variance of R_k which equals

$$Var(Y_k - A_k\theta_k|\psi_0,\tau_0,Y_k).$$

This is the same as, $Var(-A_k\theta_k|\psi_0,\tau_0,Y_k)$ which equals,

$$A_kD^*_{\theta k}A_k' .$$

The variance matrix, $D^*_{\theta k}$, is the portion of the posterior variance matrix, D^*_θ, which is pertinent only to the outcome, Y_k.

The expression can be put in a computationally more convenient form if we note that $Tr\{A_kD^*_{\theta k}A_k\}'$ can be permutated to yield,

$$TR(A_kD^*_{\theta k}A_k') = TR(A_k'A_kD^*_{\theta k}) .$$

Thus the conditional expected sufficient statistic for σ^2_k can be expressed as,

$$E(R_kR_k'|\psi_0,\tau_0,Y) = (Y_k - A_k\theta_k)'(Y - A_k\theta_k) + TR(A_k'A_kD_k{}^*_{\theta k}) .$$

This can be translated back to the mixed model form to yield an equation similar to what was used in the multilevel path program,

$$(Y_k - A_{1k}\theta_{1k} - A_{2k}\theta_{2k})'(Y_k - A_{1k}\theta_{1k} - A_{2k}\theta_{2k}) + TR \begin{vmatrix} \bar{A}_1'A_1 & A_1'\bar{A}_2 \\ \_A_2'A_1 & A_2'A_2\_ \end{vmatrix} \begin{vmatrix} \bar{D}^*_{11} & D^*_{12}\bar{} \\ \_D^*_{21} & D^*_{22}\_ \end{vmatrix} .$$

This involves rather large matrices. By multiplying the partitioned matrices, expanding and simplifying terms this can be broken down to a tractable computational formula in terms of group-level variables.

As with the estimation of τ, the maximization step is relatively trivial. The estimated sufficient statistics is simply divided by N to yield the maximum likelihood estimate for σ^2_k. The above steps are repeated for all K of the σ^2 terms and the diagonal terms of ψ are constructed from these K estimates.

At this point we have τ_1 and ψ_1 for one iteration of the EM algorithm. These variance matrices are then used to calculate all of the terms in the likelihood expression from Equation 2.40,

$$-\text{Log}|\Psi| - \text{Log}|T| + \text{Log}|D_{11}| + \text{Log}|C^{-1}| - Y\Psi^{-1}(Y - A_1\theta_1^* - A_2\theta_2^*).$$

This is the criterion for convergence. The change in the likelihood is positive from each iteration to the next as the likelihood increases towards a maximum. If the positive

change in likelihood is less than .01%, then it is judged that the algorithm has converged to the maximum likelihood estimates.

III. Test Statistics

Three test statistics are utilized in the program. Two provide tests for variances and one for second-stage regression coefficients.

Statistical Test for Parameter Variances

Recall that the second-stage model for the paths for group j is,

$$B_j = W_j\gamma + U_j \ .$$

The variance of U_j is τ, which is a P by P variance/covariance matrix. The P diagonal elements of τ, τ_{pp}, are the parameter variances of the paths. The larger this variance is the more the structural parameter varies from group to group. If the parameter variance is zero, the corresponding path is considered to be the same for all groups, and is therefore a fixed quantity. This has great implications for interpreting an analysis, so it is useful to have a statistical test of whether the parameter variance of a path is zero.

Such a test is a chi-square statistic which for group j consists of the ratio of the estimated total variance of the path (parameter variance + sampling variance) over the parameter value of the sampling variance. This ratio is summed over all J groups;

$$\sum_{j=1}^{J} \{\text{Total Variance of } B_p \text{ / Sampling Variance of } B_p\} .$$

Since the numerator is the sum of parameter and sampling variance, under the null hypothesis that parameter variance is zero, the ratio should be small. Conversely, assuming the parameter variance is not null, as the parameter variance gets large so does the test statistic.

A statistic which is estimable in terms of the current model is, according to Hedges (1982),

$$\sum_{j=1}^{J} (\hat{B}_{jp} - W_{jp}\gamma^*_p)^2 / V_{ppj}) ,$$

where, $\hat{B}_{jp}$ is the least squares estimate for path p for group j,

W_{jp} is the second-stage predictor matrix for path p and group j,

γ^*_p is the posterior estimate of the s by 1, second-stage regression coefficient vector for path p,

V^*_{ppj} is the sampling variance for path p. Its estimate consists of the p,p element from the matrix, $(Z_j'\psi^{-1}Z_j)^{-1}$, where Z_j is the first stage predictor matrix and ψ is the estimated variance of first -stage errors. This is the familiar least squares estimate for sampling variance of a regression weight.

This statistic has an asymptotic chi-square distribution with J-S degrees of freedom with J equal the number of groups and S equal the number of second-stage predictors for path p.

Note that for this to be a true chi-square test it is assumed variance term for each group, V_{ppj}, is a known

parameter. Raudenbush & Bryk (1986) point out that since the sample size over all groups is usually large, this is not a hazardous assumption. They point out a more serious problem with the statistic, though. It may be sensitive to departures in normality of Y and U. Raudenbush and Bryk suggest that this statistic should be interpreted with caution unless the probability is very small, e.g. in the .001 range.

The Percent of Variance Accounted For by the Second-Stage Model

As we will see in the analysis chapter, two models are compared in a multilevel analysis, an unstructured between-group model and a structured between-group model. The unstructured model stipulates that first-stage paths vary about a grand mean path;

$$Y = ZB + R$$

$$B = B\text{-Mean} + U;$$

where, B-Mean is the vector of grand mean paths. The motive behind running this model is to get an estimate of the total parameter variance of the paths, unconditional an a between-group regression.

Typically we would proceed to specify a structured between-group model in a subsequent run;

$$Y = ZB + R$$

$$B = W\gamma + U,$$

where the second-stage intercepts are incorporated into γ,

W is the matrix of between-group predictors, and

γ is the vector of regression coefficients for paths.

The variance of U_j, (τ) now represents the residual variance matrix of the second-stage model, conditional on the second-stage predictors. In the ideal case where $W\gamma$ predicts perfectly, the second-stage model would account for 100% of the parameter variance and τ would be zero. A simple test for the percent of variance accounted for by the between-group model is given by Raudenbush and Bryk (1986);

$$\{Var(B_{jp}) - Var(B_{jp}|Wp)\} \; / \; Var(B_{jp}) \; ,$$

where, $Var(B_{jp})$ is the unconditional parameter variance of path p. This is the p,p element from the τ matrix estimated in the unstructured between-group model,

$Var(B_{jp}|W_p)$ is the conditional variance of path p. This is the parameter variance estimated in the structured between-group model.

This statistic provides a useful criterion by which to assess overall model performance and it also provides information for modifying the between-group model for each path.

Z-Test for Second-Stage Regression Coefficients

An ordinary Z statistic can be used to test whether a second-stage regression coefficient is zero. The standard Z form is;

$$\frac{\hat{P} - P}{\text{Standard Error}(\hat{P} - P)}$$

where; $\hat{P}$ is the estimated parameter value,

P is the hypothesized parameter value,

Standard Error $(\hat{P} - P)$ is the estimated standard error of the difference score between the estimate and the hypothesized parameter value.

In terms of the multilevel path analysis this becomes,

$$\frac{\gamma^*_a}{D_1^{*1/2}{}_{s,s}}$$

where, γ^*_s is the second-stage parameter coefficient, and

$D_1^{*1/2}{}_{s,s}$ is the square root of the s,s diagonal term from the posterior sampling variance of the fixed effect, i.e. from $D^*_{\theta 1}$ in Equation 2.22.

This statistic has an asymptotic Z distribution.

Raudenbush & Bryk (1986) contended that statistical tests of regression coefficients would be more robust to such violations than chi-square tests of variances.

A note of caution has to do with the sheer number of Z-tests that can occur. Each path can have numerous between-group predictors so we could find ourselves performing myriad non-independent Z-tests. The overall alpha level of the entire set of Z-tests is unknown. It is therefore advised that these tests only be use as a rule-of-thumb and not as proof of the existence or nonexistence of particular effects.

IV. Accuracy Check of the Computer Path Algorithm

The computer program to perform the multilevel path analysis involves thousands of calculations over numerous iterations of the EM algorithm. Checking the accuracy with which the statistical formulae were translated into code by hand calculations would be an unwieldy task, and would be quite prone to error. It was therefore concluded that the only reliably accurate way to check the equations and the design of the program would be to compare it to an already established estimating program.

The multilevel path model is an elaboration of the Hierarchical Linear Model devised by Raudenbush (1984), so the HLM program which estimates this model (Bryk, et al, 1986) is a natural choice for comparison. The difference between the two models is that the multilevel path model stipulates a multiple equation system, without intercepts at the first-stage; while the HLM model stipulates a regression model with intercepts. I therefore modified the multilevel path program so that the first-level design could include intercepts and could be restricted to one equation. With these modifications the models for the two programs should be the same. Also, under these conditions the Bayesian estimating equations of the multilevel path model should reduce to the HLM case.

To empirically test whether the algorithms were equivalent

in this case, both programs were used to analyze an identical dataset stipulating an identical model for the data.

The Model and Data

The data to be given the parallel analysis was drawn from the High School and Beyond study (Coleman, Hoffer & Kilgore, 1982). This study will be more fully described in chapter four. For the purposes of exposition I will only mention that in this analysis the data consisted of measures on students in 94 schools. The dataset also included measures at the school level but they were not included in the validation run.

The within-school model is a standard regression with one outcome and three predictors. For individual i and group j this model is,

$$\text{Math Achievement}_{ij} = B_{0j} + B_{1j}(\text{Minority Status})_{ij} + B_{2j}(\text{Gender})_{ij} + B_{3j}(\text{SES})_{ij} + R_{ij}$$

where;

Math Achievement = a standardized math score,

B_{0j} = is the mean math achievement for school j. The student level predictors were mean deviated so that the intercept was the group mean.

Minority Status = Whether the student was a minority,

Gender = Whether the student was male or female,

SES = Socioeconomic status index for student,

R_{ij} = Sampling error.

The between-group model was unstructured,

$$B_j = \text{B-Mean} + U_j ,$$

where, B_j = The 4 by 1 vector of regression coefficients for

a group,

B-Mean = The 4 by 1 vector of grand mean paths,

U_j = The 4 by 1 vector of parameter errors.

This model was run on both analysis programs for 20 iterations of the EM algorithm.

The degree of agreement was quite high. The likelihood function used to monitor the progress of the algorithm, Equation 2.22, incorporates all the information of the estimates. The value of the likelihood functions for the two programs differed only from .001% to .02% over the 20 iterations. The estimates for the first-stage error variance were identical at 1.498. Estimates for the elements of τ were very close, with differences ranging from .002% to 11% (see tables 5.1 and 5.2). The posterior estimates for the second-stage intercepts, γ_0, were also very much in agreement, with differences ranging from 0% to .07% (see table 5.3).

The two estimating programs produced virtually identical parameter estimates when identical models were analyzed. What little differences there were can be explained by the fact that the programs were not written in the same programming language and the form of the equations were not the same. The HLM program was written in fortran with self-contained matrix subroutines. The multilevel path model, on the other hand, was written in SAS, using the Proc Matrix procedure (SAS Institute, 1985). Rounding errors and the accuracy of subroutines could differ between the two languages.

This parallel analysis has established that the multilevel path program produces sensible results when compared to an established estimating program for a restricted model. The soundness of the algorithm has not been established for other models and other datasets. The program will have to be run under many conditions before the status of programming bugs can be thoroughly assessed.

CHAPTER IV:

USING THE MODEL

In chapter two the multilevel path model was defined, the estimators derived and their statistical properties discussed. In this chapter we ask if this model can be fruitfully applied to educational research. We need to know a) if the model gives estimates which have some meaningful correspondence to the actual processes being studied, and b) if a multilevel path analysis lends itself to an interpretation which enhances our understanding about important educational questions. In order to demonstrate the meaningfulness and interpretability of the model I will analyze two educational data sets.

I. The Analysis of the High School and Beyond Data

The first data set is from the High School and Beyond study (Coleman, et al, 1982). As mentioned in chapter one this was a very large scale study in which a sample of 998 was taken nationwide. A major focus of the study was the question, "What is the relationship between students' characteristics, such as family background and ethnicity, and academic success?" Previous studies have shown that the relationship between students' SES and academic attainment is substantial (Lee, 1986). This finding is of great concern to many educators because it seems to undercut the ideal of fairness, equality and equal access to opportunity. The study by Coleman et al (1981) focused on the relationship

between student background and achievement in high schools of two types, public and private Catholic. In this study, background and achievement were examined in a broad context. Within the schools, student background information was gathered on variables such as family SES and student's ethnicity. Academic measures included number of math classes taken and mathematics achievement score on a standardized math test. One of the most controversial conclusions of this study was that Catholic schools were found to be more egalitarian than public schools. This claim was made on the basis of the finding that the relationship between SES and achievement is not as strong in Catholic schools as it is in public schools, so that the disequalizing effect of SES on educational outcomes is smaller in the Catholic sector. This has led to widespread speculation that not only is much education inequitable in this country, but that such inequity is concentrated in our public institutions.

On the face of it, the appearance of inequity in public schools is alarming and invites speculation about "What is wrong with our schools?" In order to gauge the seriousness of the problem and to devise solutions, the mechanism behind the inequity must be understood.

Multilevel linear models are particularly well suited for exploring this question because the issue concerns processes that arise at different levels of aggregation. The effect of SES on achievement pertains to students within schools. The influence that sector has on the SES to achievement

(SES—>Ach) effect pertains to a school-level variable (sector) and its effect on a student-level process. The mechanism which explains how sector influences the SES——>Ach effect would consist in school-level variables which characterize public and Catholic schools and explain why the two types of schools function differently. For example, it may be discovered that certain policies and practices characterize Catholic schools and explain why students of different SES backgrounds achieve at the same level in these schools.

One effort to bring a multilevel approach to bear on this issue was the reanalysis of Coleman, et al's study by Raudenbush and Bryk (1986). They used an approach called the Hierarchical Linear Model, or HLM, in which a single-outcome, multiple regression is posited in numerous groups as the within-group model. The variation in the group parameters over groups is modeled at the between-group model in such a way that characteristics of the group predict the group's regression coefficients. Raudenbush and Bryk demonstrated the existence of inequity within schools by showing that there are schools which have a positive SES—>Ach regression slope. Further, they demonstrated that public schools were less equitable than Catholic schools by showing that being in the public sector was positively associated with a school having a larger SES——>Ach slope. In other words, sector was introduced as a predictor in the between-group model. Lee (1986) and Lee and Bryk (1986) carried this logic further. Their goal was to demonstrate that

there was a mechanism which explained the greater equity of Catholic schools. They added variables to the between-group model which represented policies and practices of schools. If the proper explanatory policy variables were introduced, the estimated effect of sector on the SES——>Ach slope would disappear.

It is noteworthy that Raudenbush and Bryk and Lee and Bryk upheld the existence of a sector effect because Coleman et al. estimated this effect by performing separate student-level regressions first for the public school students and then for the Catholic school students. The classroom level of analysis was ignored, making the results vulnerable to aggregation bias.

Another possible source of bias is the presence of confounding variables at both the within-class and the class levels. Raudenbush and Bryk devised a model that controlled for confounding variables at both levels of analysis. They concluded that there remained a sector effect on the SES/achievement relationship, even when aggregation bias and confounding factors were controlled for.

In Lee and Bryk's analysis two outcomes were used as yardsticks of academic attainment: the number of math courses taken in high school and math achievement. In the HLM approach this means that two separate within-class models must be analyzed. One had math achievement as the dependent variable, predicted by student background. The other had number of

math classes as the dependent variable, also predicted by student background.

The first within-class model had the form:

$$\text{Ach} = B_0 + B_1(\text{Academic Background}) + B_2(\text{Minority}) + B_3(\text{SES})$$

In the second-stage model, the first-stage slopes, B_2 (Minority-->Ach) and B_3 (SES-->Ach) are predicted by school context variables (i.e. average school SES and percent minority enrollment in school) and school practice and climate variables (e.g. number of math courses available in the school and level of disciplinary problems in the school). The parameters, B_0, the school mean, or B_1, serve as statistical controls in this analysis so the discussion here focuses on B_2 and B_3. After school context and school climate variables were taken into account in the second-stage model, the sector effect disappeared. So the notion of "inequity" is explained away and is replaced by the specific climate, policies and practices of the school.

In the second model, the number of math classes is the outcome for the within-class regression:

$$\#\ \text{Classes} = B_0 + B_1(\text{Academic Background}) + B_2(\text{Minority}) + B_3(\text{SES})\ .$$

In the subsequent analysis, the between-group predictors for the B_2 (Minority-->Classes) slope and the B_3

(SES-->Classes) slope were school climate variables (i.e. minority enrollment and school SES). This analysis was much less conclusive than the previous one. The sector effect was never explained away.

A limitation of the Raudenbush and Bryk analysis is that it only modeled one outcome. In Lee's analysis it is contended that the two outcomes, number of classes taken and achievement, are both important outcomes which bear on the equity issue. But the main limitation of Lee's approach is that the two outcomes of interest must be assessed by two separate analyses. AS Lee asserts (1986), it is reasonable to assume that the number of math classes taken has a strong effect on math achievement. This implies that a properly specified model would include the effect of Number of Classes on achievement. Such an analysis requires path modeling and is outside the scope of the HLM model. The analysis presented in this chapter employs this sort of within-groups path model.

The issue can be illustrated by path diagrams. A model similar to the two parallel within-groups regression models used by Lee would have the form:

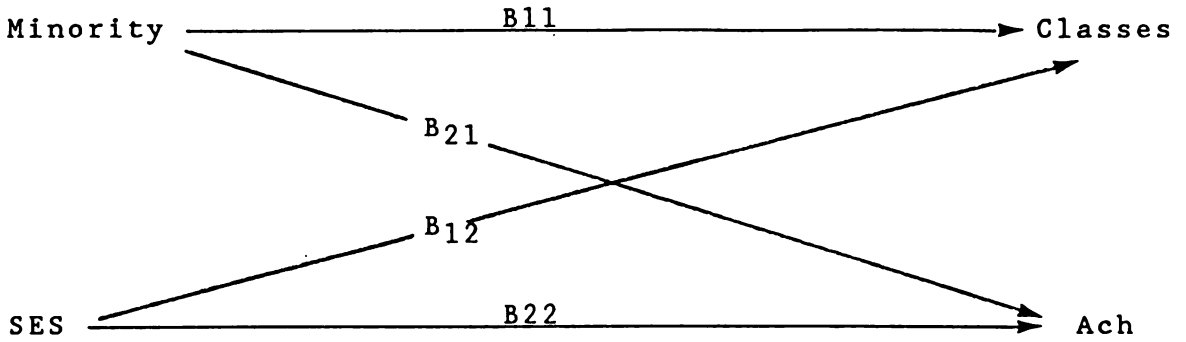


Figure 4.1
Path Diagram of Two Separate Regression Analyses

A separate regression analysis is run for each outcome. The slope estimates of one analysis does not effect the slope estimates of the other analysis. But what if we connect the separate models by drawing an arrow between the outcomes:

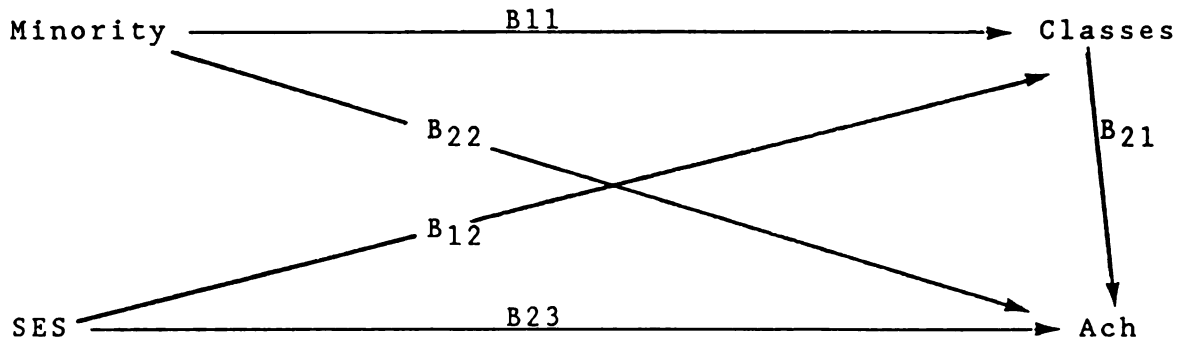


Figure 4.2
A Single Path Model Incorporating All Variables

Instead of two regression models we have one path model. Such a model is different from separate regression models in two ways 1) an additional relationship, the Classes/Ach effect, is estimated and 2) some of the previous relationships may be estimated to have very different values under this model. For

example, suppose minority status and student SES affect achievement only by affecting the number of classes taken, i.e. suppose the actual model is;

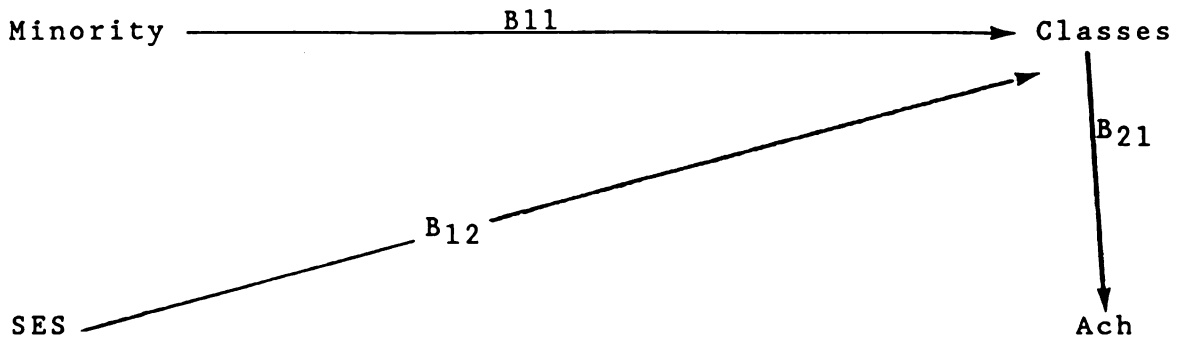


Figure 4.3
Path Model With Indirect Effect of
Student Background on Achievement

If this represents the state of affairs of the world, when the Classes/Ach path is added to the model, the estimates of Minority/Ach and the SES/Ach path will tend towards zero. Contingencies such as these can only be explored by a path analysis.

The sample and the Data

In the analysis I performed on the High School and Beyond data a random sample of 158 schools out of the original base of 998 schools was employed. In this sample there were 68 Catholic schools and 90 public schools. Four within-group variables were used in the present analysis:

Minority Status (Minority) - Whether or not the student was a minority. 0-White, 1-Minority.

SES - A composite index of student's social class.

Number of Classes (Classes)- Number of advanced math classes taken in high school.

Math Achievement (Ach) - Senior year math achievement.

Two Parallel Within-Class Regression Models

In order to demonstrate the explanatory power of a within-groups path analysis, we will first estimate the two parallel but separate regression models illustrated by figure 4.1. We will then compare these results to an analysis employing a within-group path analysis as depicted by figure 4.2.

The two regression models have the single-equation form for person i and group j,

$$Classes_{ij} = B_{11j}(minority) + B_{12j}(SES) + R_{ij1}$$

$$Achievement_{ij} = B_{21j}(Minority) + B_{22j}(SES) + R_{ij2}$$

The between-groups model is unstructured, i.e. there are no group-level predictors so that regression parameters vary about a grand mean,

The results of the parallel regression run can be found in table 1. In table 1-A, the estimated parameter variances are listed. For each estimated parameter variance there is a corresponding chi-square test. This chi-square statistic tests the null hypothesis that the parameter variance is zero (see chapter 3). A larger chi-square statistic indicates a less probable result given the null hypothesis. If the probability of the chi-square test is below some a priori

critical level, that is grounds for inferring that the parameter variance is not zero. The exact value of the critical probability is of course arbitrary, but the customary .05 value will be assumed.

In table 1-A the chi-square tests indicate that all the parameter variances are significant. In other words every regression coefficient varies from group to group.

Table 1-B shows the weighted average of the coefficients, with the coefficients from each group weighted by the precision of the group estimate. This gives us some idea of an 'average' regression model from which all the groups deviate. In the special case in which the coefficient is inferred to have zero parameter variance the average coefficient represents the structural relationship for all groups. The Z test indicates whether the average coefficient is significantly different from zero. As we see, all coefficients are different from zero.

This analysis is dramatically more interesting when compared to the within-group path model in the next section.

Table 1-A

High School and Beyond Data

Parameter Variance of Paths

Parallel Regression Model

Path	Parameter Variance	Chi-Square Statistic	Probability	Parameter To Total Variance
B_{11}	.324	232.616	<.0001	.285
B_{12}	.059	219.700	<.0001	.235
B_{21}	6.692	211.456	<.0001	.270
B_{22}	.688	160.966	.034	.156

Table 1-B

High School and Beyond Data

Average Value of Paths

Parallel Regression Model

Path	Average Value	Z Statistic	Probability
B_{11} (Minority-> Classes)	-.279	-3.620	.0003
B_{12} (SES-> Classes)	.343	10.420	<.0001
B_{21} (Minority-> Ach)	-2.690	-7.318	<.0001
B_{22} (SES-> Ach)	1.326	9.189	<.0001

The Within-Group Path Model

In contrast to the parallel regression analyses we now propose a path model which 1) relates background variables to both Classes and achievement, but also relates Classes to achievement as depicted by figure 4.2. The within-group path model is a two equation system which has the individual form (for individual i and group j):

$$Classes_{ij} = B_{11j}(minority) + B_{12j}(SES) + R_{ij1}$$

$$Achievement_{ij} = B_{21j}(Classes) + B_{22j}(Minority) + B_{23j}(SES) + R_{ij2}$$

Unconditional Between-Group Analysis

The first computer run that is performed with a multilevel path analysis specifies an 'unconditional' between-group model, one that stipulates no between-group predictors, as was the case with the parallel regression analyses. Since the parameter variance estimate is not conditioned on between-group predictors it is at its maximum possible value. Unconditional estimates of parameter variance provide baseline estimates of the total variance, if any, that may be explained by future runs which include group-level variables. This baseline run will also yield the mean slopes across groups, giving us an idea of what the central tendencies of the paths are. The unconditional second-stage model takes the simple form:

$B_{11} = \bar{B}_{11} + U_{11}$

$B_{12} = \bar{B}_{12} + U_{12}$

$B_{21} = \bar{B}_{21} + U_{21}$

$B_{22} = \bar{B}_{22} + U_{22}$

$B_{23} = \bar{B}_{23} + U_{23}$

The $\bar{B}_{kp}$ terms are the average, or pooled-within-group, estimates of the paths. If there is no parameter variance, i.e. if $Var(U_{kp})=0$, then the pooled within group path is characteristic of all groups. Figure 4.4 is a multilevel path diagram of the baseline model. The bold arrows represent the first level (within-group) paths. The finely etched arrows represent predictive relationships between the second level (between-group) variables and the first level paths. In this model only the second-stage errors (U_{kp}) impinge on the first-stage paths.

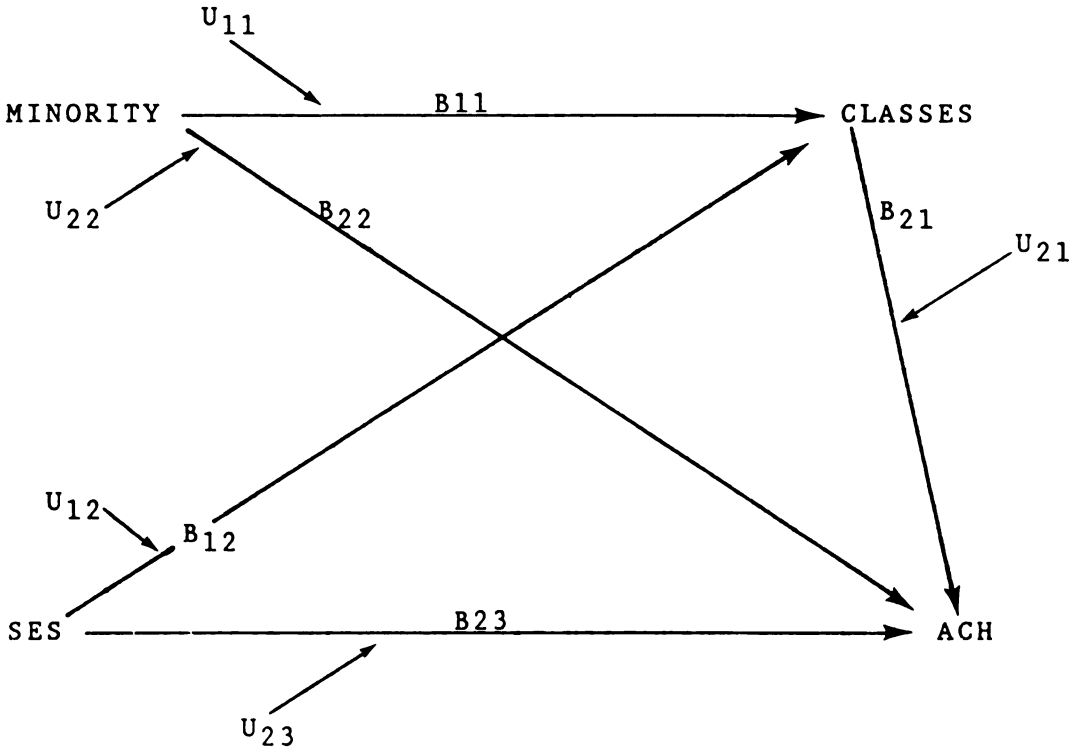


Figure 4.4
High School and Beyond Data: Baseline Model

We can see from table 1-C that parameter variance is only a small part of the total variance of the path estimates. Column five lists the ratio of the parameter variance to the total variance. The numerator is from column one, the parameter variance. The denominator is a statistic which represents the sum of parameter and sampling variance. This ratio, then, is the percentage of total variance represented by the parameter variance. The estimated parameter variance of the first three paths accounts for only 28%, 24%, and 22% of total variance, respectively. The relatively small group sizes could account for why the sampling variance is large when compared to parameter variance.

Column four lists the probabilities of the chi-square tests. From this we see that the first three parameter variances are significantly different from zero. The parameter variance for B_{22} represents only 13% of the total variance, even though the chi-square test still indicates that this is a non-zero quantity (p=.007). The last path, B_{23}, has a parameter variance which is only 9% of total, and indeed the chi-square indicates this is not significantly different from zero (p=.95). With virtually no systematic variance to be explained, it is unlikely that any between-group variables will predict the B_{23} path.

Table 1-C

High School and Beyond Data

Parameter Variance of Paths

Unstructured Between-Group Model

Path	Parameter Variance	Chi-Square Statistic	Probability	Parameter To Total Variance
B_{11}	.319	322.601	<.0001	.281
B_{12}	.059	219.546	<.0001	.236
B_{21}	.335	210.709	<.0001	.221
B_{22}	2.173	172.674	.0073	.130
B_{23}	.198	104.037	.955	.087

This model offers a sharp contrast to the parallel regression analyses. First of all let us compare the average coefficients from the two analyses in tables 1-B and 1-D. The first two coefficients, Minority/Classes and SES/Classes, arevirtually unchanged. But the last two coefficients, Minority/Ach and SES/Ach, have become much smaller. The average minority/Ach effect went from -2.69 to -1.928, while the average SES/Ach effect went from 1.326 to .391. From this we can conclude that on the average, much of the effect of students' background on achievement is through the number of classes taken. In other words, student background determines achievement largely by determining how many classes the student will take.

It might be argued that we have set up a straw man, that the HLM approach could have modeled the same two equations as the multilevel path model, in two separate runs. This is a viable option but it allows for less satisfactory modeling at the between-group stage. With the multilevel path analysis the paths from all equations can be modeled by group variables as one set. Since the covariances among paths across equations are accounted for, the simultaneous approach can be expected to yield more appropriate results for the second-stage model.

Table 1-D

High School and Beyond Data

Average Value of Paths

Unstructured Between-Group Model

Path	Average Value	Z Statistic	Probability
B_{11} (Minority-> Classes)	-.278	-3.620	.0003
B_{12} (SES-> Classes)	.342	10.414	<.0001
B_{21} (Classes-> Ach)	2.951	38.029	<.0001
B_{22} (Minority-> Ach)	-1.928	-7.411	<.0001
B_{23} (SES-> Ach)	.391	3.622	.0003

Another striking contrast occurs if we compare the regression model parameter variances, table 1-A, with the path model parameter variances, table 1-C. As before, the values for the first two coefficients are virtually unchanged between the two models. But the variances for the last two coefficients have diminished by two-thirds from the regression to the path models. In fact, the path model variance for the SES/Ach effect is not significantly different from zero. The SES/Ach effect is constant over groups when achievement is controlled for the number of classes taken. Almost all of the variation in equity is accounted for by variation in how classes are distributed. The situation is like the path model depicted in figure 4.3 where SES affects achievement through Classes. In order to explain apparent differences in the SES/Ach relationship from school to school, we must find school-level variables which explain the SES/Classes effect and the Classes/Ach effect. This is the issue taken up in the structured between-group analysis.

Structured Between-Group Analysis

A second statistical analysis is now presented in which group-level predictors have been included in the second-stage model. The three group-level variables that were used in this analysis are defined as follows:

Sector	- Whether the school belonged to the public school or the Catholic sector. 0=Public, 1=Catholic.
Ave-SES	- Average social class of all students in the school.
Sd-SES	- Standard deviation of students' social class in a school.

Several models were estimated to explore different combinations of second-stage predictors. The multilevel path analysis is highly sensitive to changes in the model. Because second-stage predictors can be multicollinear and because the estimation procedure is full information maximum likelihood, the estimate for one parameter affects estimates for all other parameters. Note that the first-stage model does not change. After some exploration a second-stage model of the following form was settled upon:

$$B_{11}(\text{minority-->classes}) = \bar{B}_{11} + \gamma_{111}(\text{sector}) + U_{11}$$

$$B_{12}(\text{SES-->Classes}) = \bar{B}_{12} + \gamma_{121}(\text{Sector}) + \gamma_{122}(\text{Ave-SES}) + U_{12}$$

$$B_{21}(\text{Classes-->Ach}) = \bar{B}_{21} + \gamma_{211}(\text{Sector}) + U_{21}$$

$$B_{22}(\text{Minority-->Ach}) = \bar{B}_{22} + \gamma_{221}(\text{Sd-SES}) + U_{22}$$

$$B_{23}(\text{SES-->Ach}) = \bar{B}_{23} + U_{23}$$

As with the baseline analysis, the intercepts of the between-group regression are slope averages. This is because all between-group predictors were mean deviated. A pictorial

depiction of this multilevel path model is given by figure 4.5. As with figure 4.4, the bold arrows represent the first-level model and the finely etched arrows represent a predictive relationship between group-level variables and paths. As before, the U's are parameter errors.

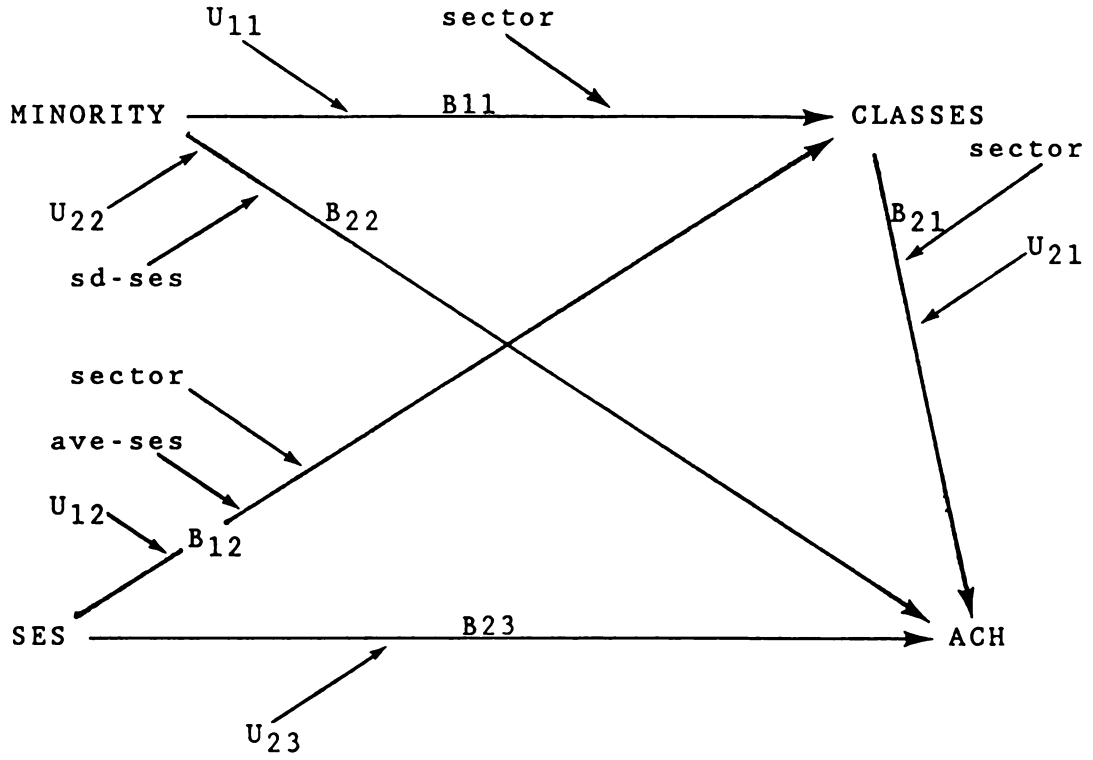


Figure 4.5
High School and Beyond Data: Baseline Model

The striking feature of this model is that sector does not enter into the relationship between student's background (i.e. minority status and SES) and achievement. Raudenbush & Bryk (1986) and Lee (1986) both found such relationships. We see that sector helps determine B_{11}, the Minority-->Classes path, B_{12}, the SES-->Classes path and B_{21}, the Classes-->Achievement path. The implication would seem to be that once the effect of Classes on Achievement is taken into account for students within schools, sector no longer determines the relationship between background and achievement. Such a conclusion would not be apparent without a path model at the within-group level.

Sector <u>is</u> important for mediating the relationship between background and number of classes. The conclusion that we draw is that sector mediates equity but <u>not</u> by directly influencing the relationship between students' background and achievement. Rather, sector influences the indirect relationship between background and achievement. In other words, Catholic schools accomplish greater equity in achievement by promoting equity in classes. This is demonstrated by the sector influence on the Minority/Classes path and on the Classes/Ach path.

Lee found that this relationship was resistent to being explained away by school context and school climate factors. Thus a mechanism which explains how Catholic schools function differently from public schools, was not found.

Table 2-A gives information about the parameter variances and helps us to assess how well the between-group model fit

the data. If between-group variables in the model perfectly predicted a path, the parameter variance would fall to zero. Although no model achieves this ideal the last column of table 2-A, the R^2 column, helps us assess how close the model came to the ideal. R^2 is the proportion of parameter variance accounted for when compared with the baseline analysis (Raudenbush & Bryk, 1986). The formula is:

$$R^2 = \frac{Var(B) - Var(B|W)}{Var(B)}$$

Var(B) is the parameter variance for a path from the baseline model. Var(B|W) is the parameter variance from a model in which the parameter variance is conditioned on group level predictors, W. The R^2 for B_{11} is .32, so Sector accounts for 32% of the total parameter variance for this path.

The estimated second-stage regression coefficients are given in table 2-C. In column three we find that the coefficient for Sector predicting the B_{11} path is .62. The Z test of whether this coefficient is different from zero has a probability less than .0001, which is convincingly significant.

The test of the usefulness of the model is in what it can say about school processes. From table 2-B we see that the average value of B_{11} is -34, which is significant at probability < .0001. This indicates that on the average being a minority leads to having fewer math classes. As we

have seen, the second-stage regression coefficient of Sector predicting B_{11} is .62. This means that going to a Catholic school will have a positive effect on the B_{11} path (i.e. makes the slope less negative by an increment of .62). The effect of being in a Catholic school (W=1) is demonstrated by the second-stage regression equation:

$$\hat{B}_{11} = -.34 + (1)\ .62$$

Being in a Catholic school flips the sign of the Minority/Classes path from -.34 to; -.34 + .62 = .28. In public schools being a minority is a disadvantage for taking math classes, while in Catholic schools it is an advantage. This defines a disordinal interaction between sector and the Minority-->Classes path.

Now let us focus attention on the other background variable, SES. Table 2-B indicates that the average SES-->Classes path is .35. On average, higher SES students take more math classes. This B_{12} path is predicted by two group level variables:

1) Sector, which has the regression coefficient of -.25. In Catholic schools the B_{12} path is smaller indicating that there is a weaker relationship between SES and number of classes taken. Catholic schools seem more egalitarian by this criterion.

2) Ave-SES has a .15 coefficient for predicting B_{12}. In schools with higher average SES, the student's SES is more important for determining number of classes taken, i.e higher

SES schools are less egalitarian.

Table 2-A

High School and Beyond Data

Parameter Variance of Paths

Structured Between-Group Model

Path	Variance	Parameter Statistic	Chi-Square Probability	Parameter To Total Variance	R^2
B_{11}	.216	204.312	<.0001	.203	.32
B_{12}	.048	205.549	<.0001	.198	.19
B_{21}	.318	207.220	<.0001	.213	.05
B_{22}	1.849	166.351	.0149	.109	.15
B_{23}	.187	103.918	.955	.082	——

Table 2-B

High School and Beyond Data

Average Value of Paths

Structured Between-Group Model

Path	Average Value	Z Statistic	Probability
B_{11} (Minority-> Classes)	-.343	-4.705	<.0001
B_{12} (SES-> Classes)	.353	11.010	<.0001
B_{21} (Classes-> Ach)	2.953	38.323	<.0001
B_{22} (Minority-> Ach)	-1.868	-7.319	<.0001
B_{23} (SES->Ach)	.371	3.451	.0006

Table 2-C

High School and Beyond Data

Second Stage Regression Coefficients

Path Predicted	Second Stage Predictor	Second Stage Regression Coefficient	Z Statistic	Probability
B_{11}				
	Sector	.617	4.34	<.0001
B_{12}				
	Sector	-.246	-3.52	.0004
	Ave-SES	.149	1.80	.0722
B_{21}				
	Sector	.278	1.82	.069
B_{22}				
	Sd-SES	-7.02	-2.72	.007
B_{23}	———	———	———	———

The direction of all the second-stage effects is coincident with previous research and substantive theory. The R^2 for the B_{12} slope indicates that these three second-stage predictors (Ave-SES, Sector and Ave-Classes), account for 22% of the total parameter variance of the path.

The B_{21} path, representing the relationship of the number of math classes a student takes to math achievement, has an average value of 2.95 (as indicated by table 2-B). Taking more classes is strongly related to higher achievement. This relationship is quite variable across schools, with an easily significant parameter variance indicated in table 2-A. It is helpful to look at the "parameter to total variance" ratio in column 4. Twenty one percent of the total variance is parameter variance even after being conditioned on the second-stage model.

The single between-group variable that predicts the B_{21} (Classes-->Ach) path is Sector. The regression coefficient of B_{21} on Sector is .28 (re table 2-C) which is only marginally significant (P = .07). The interpretation that can be given this is that Catholic schools evidence a somewhat stronger positive relationship between number of classes and achievement than public schools. It could be said that classes are more efficient in Catholic schools, i.e. taking a math class creates a greater gain in math achievement in Catholic schools. To find a mechanism for this influence we might inquire into curricular differences between Catholic and private schools.

As table 2-A shows, the R^2 for B_{21} is only .05, i.e. only 5% of the parameter variance is explained by Sector. Sector is not very important for mediating the Classes to Ach effect, and there are other factors, not represented in this analysis, which would explain the path.

The final two paths represent the relationship between background variables and achievement. Looking at table 2-A we see that the parameter variance of both paths is a small percentage of total, account for only 11% and 8% of total variance. Once the number of classes is controlled for there is little variation from school to school in the relationship between student's background and achievement.

The Minority to Achievement path, B_{22}, has an average value of -1.87 (table 2-B), indicating that being a minority has a negative effect on achievement. This path is predicted by the standard deviation of SES for a school (Sd-SES). From table 2-C we see that the coefficient for Sd-SES is -7.0 (significant at a P = .007). This second-stage coefficient implies that as a school gets more heterogenous in its social mix, a student's minority status is a bigger determinant of achievement. The precise interpretation to give this relationship in terms of school processes would be difficult to determine without more information about how the schools functioned. The last column of table 2-A indicates that Sd-SES accounted for only 15% of the parameter variance in B_{22} (Minority-->Ach). Given that the total parameter

variance for B_{22} was initially quite small, the import of the Sd-SES prediction is minimal.

The final path is SES to achievement, B_{23}. It has an average value of .37, which is significantly different from zero (p=.0006, from table 2-B). Since there is virtually no parameter variance in this path, the average value represents the relationship for every school. As with number of classes taken, higher SES is associated with higher achievement scores. This holds equally true for the public and the Catholic sector.

The multilevel path model indicates that the processes in the schools that are responsible for making the Catholic seem more 'egalitarian' than public schools pertain to how math classes are distributed to students. Once we account for the number of math courses students take, the relationship between student background and achievement is quite constant across schools.

II. The Analysis of Scottish Schools

The second dataset that is to be analyzed was first interpreted by Willms (1985). The dataset I have access to comes from 20 secondary schools in one administrative division in Scotland. The total number of students in the dataset is 1292, so on average 65 students were sampled per school. The original intent of gathering the data was to estimate the effectiveness of each school based on the school mean on an

achievement index. In one study, "effectiveness" was controlled for student level socioeconomic background, and student level academic background (Willms, 1987). School "effectiveness" was also controlled for school level "context" factors consisting of aggregated SES and academic background.

In the 1987 analysis by Willms, a technique devised by Longford (1985) was employed for obtaining maximum likelihood estimates of covariance components in a multilevel mixed model. Using this technique, Willms was able to estimate school mean achievement controlled for by variables at two levels of analysis, i.e. the individual student level and the school level. In the present analysis the dataset will be used for a quite different purpose than originally intended. The present analysis will a) define a path model at the within-school level, b) will ascertain if the paths vary from school to school, and c) will explore the possibility of accounting for path variability with a between-school model which incorporates school context factors as predictors. Note that school means, which were the focus of previous analyses, do not appear in the present model at all.

It will be of technical interest to see how well the multilevel path analysis performs when there is a small number of groups, 20 in this case. In contrast, in the previous analysis of the High School and Beyond data there were 158 schools. Since the Scottish data has a small set of schools taken from a contiguous geographical area, the range of variation of the within school processes might be severely

restricted. This could result in small parameter variance for paths, and attenuated second-stage regression estimates.

There are five variables which make up the within-class data set:

Education of Mother (Edmoth)	- Educational level of student's mother.
Occupation of Father (Occfath)	- Occupational status of father. A sociological index of occupational status.
Number of Siblings (Numsib)	- Number of brothers and sisters.
Verbal Reasoning Quotient (VRQ)	- A verbal IQ battery, intended to represent general academic skills.
Achievement (Ach)	- An index of measures covering the last three years of secondary school.

The first three variables are intended to measure a student's socioeconomic status. The verbal reasoning score is intended to capture the student's academic background, i.e. the academic skills the student enters secondary schools with (Willms, 1987).

The three school-level variables consist of aggregated student-level measures and represent school context. It is often believed that aggregated individual level variables represent more than simply the average impact of the individuals' values. For example, if average verbal reasoning is high, a school might have a more interesting and creative curriculum, contributing to a positive learning environment,

even for those students with low verbal reasoning skills. This is an example of a variable changing its meaning from one level of analysis to another (Burstein, 1980).

The school level variables are:

Average SES (Ave-SES)	- An average socioeconomic background score. The SES score for a student was a weighted combination of education of mother, father's occupation and number of siblings, where the weights were derived from principle components analysis (Willms, 1987).
Average Occupational Status of Father (Ave-Occfath)	- Average of the student's occupational status index.
Average Verbal Reasoning (Ave-VRQ)	- Average of the students' VRQ score.

First-Stage Model

The path model devised on this data set was a two-equation system similar to the model posited for the High School and Beyond data. Although it would have been preferable to demonstrate the multilevel path approach with a very different model, (e.g. a four equation system with numerous endogenous predictors) a certain similarity between the two sets of data constrained the choice of sensible models. The two equation system has the following form for individual i, within group j:

$$VRQ_{ij} = B_{11j}(Edmoth)_{ij} + B_{12j}(Occfath)_{ij} + B_{13j}(Numsibs)_{ij} + R_{ij1}$$

$$Ach_{ij} = B_{21j}(Edmoth)_{ij} + B22j(Occfath)_{ij} + B_{23j}(Numsibs)_{ij} + B_{24j}(VRQ) + R_{ij2}$$

The path diagram depicted in figure 4.6 is more descriptive. This is parallel to the High School and Beyond model in general definition. In both cases there are two equations in the system and in the first equation student social background variables are antecedents for academic background. In the second equation social background and academic background (as an endogenous predictor) are antecedents for achievement. The parallel is further extended by the fact that in both studies the SES and academic background variables were aggregated to the school level to define school context, but more of this when the second-stage model is described. The primary reason for the striking parallel between the two data sets is the fact that they were compiled for similar reasons, to estimate academic outcomes which are controlled for factors at two levels of aggregation. Parallel purpose led to parallel structure.

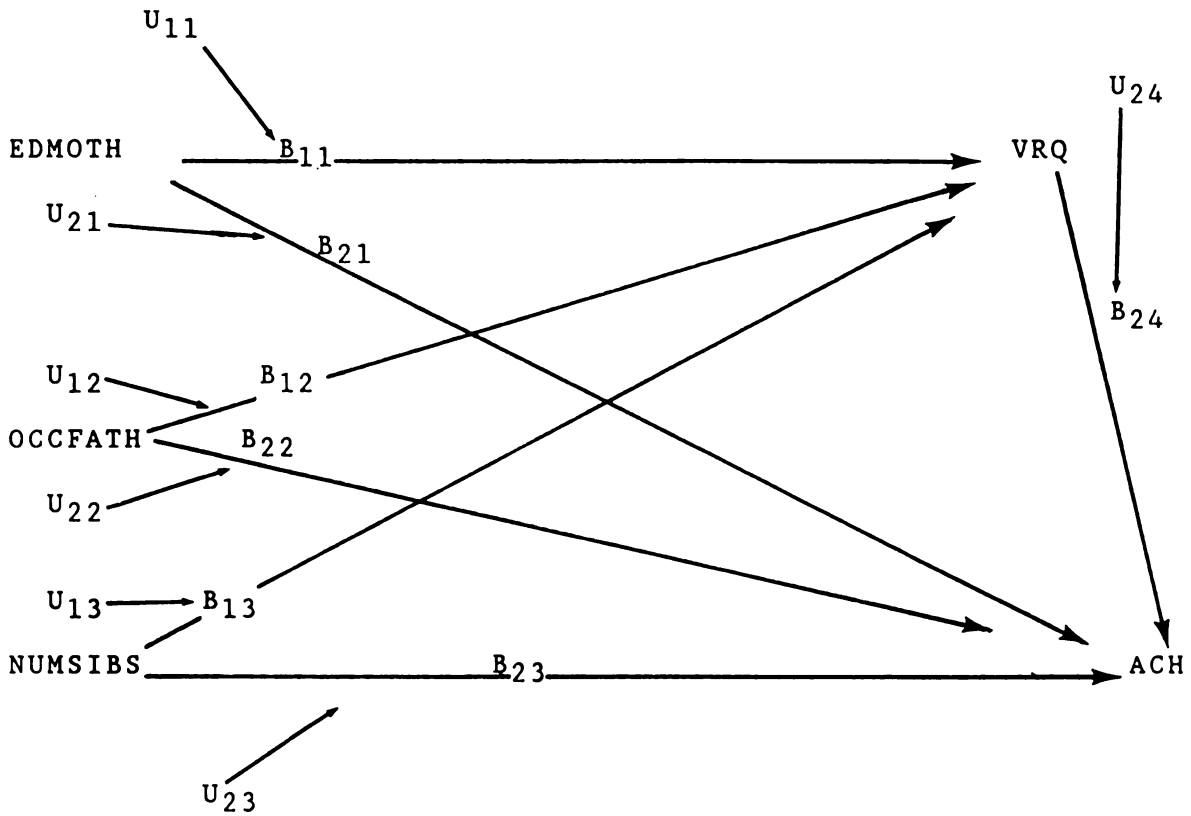


Figure 4.6
Scottish School Data: Baseline Model

Baseline Analysis

As before, an initial baseline analysis is performed with an unstructured second-stage model, i.e no school level variables are stipulated, so the paths vary around the grand mean:

$$B_{11} = \bar{B}_{11} + U_{11}$$
$$B_{12} = \bar{B}_{12} + U_{12}$$
$$B_{13} = \bar{B}_{13} + U_{13}$$
$$B_{21} = \bar{B}_{21} + U_{21}$$
$$B_{22} = \bar{B}_{22} + U_{22}$$
$$B_{23} = \bar{B}_{23} + U_{23}$$
$$B_{24} = \bar{B}_{24} + U_{24}$$

Combining this with the first-stage model gives rise to the multilevel path diagram in figure 3. As before, the bold arrows represent paths of the first-stage model, and the finely etched arrows pointing to the paths represent the impact of school level factors (in the baseline model, random parameter error) on the paths.

Table 3-A lists the estimated parameter variances of the paths. By inspecting the chi-square probabilities (column 4) it is apparent that three paths have no significant parameter variance. The paths B_{12}, B_{13} and B_{24} have chi-square probabilities of .67, .84 .34 respectively. As a result, these paths will not be modelled with school level predictors.

Table 3-B gives the 'average' betas. These are the intercepts of the second-stage regressions:

$$\bar{B}_{11}, \bar{B}_{12}, \ldots, \bar{B}_{24} .$$

The average betas coincide with commonsense expectations. B_{11}-Ave and B_{12}-Ave are positive indicating that a higher level of mother's education and a higher level of father's occupational status is associated with higher verbal reasoning skills. B_{13} is negative indicating that verbal reasoning skills are inversely related to size of family. All things being equal, having a larger family is probably associated with a generally lower socioeconomic status since more children means a greater financial burden. The same pattern of relationship between SES variables and outcome is found in the second equation. Taken together the paths, B_{21} (Edmoth-->Ach) , B_{22} (Occfath-->Ach) and B_{23} (Numsibs-->Ach), indicate that higher SES is associated with greater academic achievement. The final path, B_{24}, indicates a strong positive relationship between academic background and achievement.

Table 3-A

Scottish School Data

Parameter Variance of Paths

Unstructured Between-Group Model

Path	Parameter Variance	Chi-Square Statistic	Probability	Parameter To Total Variance
B_{11}	.02286	51.66	.0001	.60
B_{12}	.00298	16.79	.67	.23
B_{13}	.00395	13.82	.84	.39
B_{21}	.00494	33.33	.03	.35
B_{22}	.00689	35.23	.02	.45
B_{23}	.01102	45.84	.0008	.57
B_{24}	.00234	22.05	.34	.25

Table 3-B

Scottish School Data

Average Value of Paths

Unstructured Between-Group Model

Path	Average Value	Z Statistic	Probability
B_{11} (Edmoth-> VRQ)	.086	1.99	.046
B_{12} (Occfath-> VRQ)	.236	7.96	<.0001
B_{13} (Numsibs-> VRQ)	-.170	-5.72	<.0001
B_{21} (Edmoth-> Ach)	.093	3.69	.0002
B_{22} (Occfath-> Ach)	.111	4.01	<.0001
B_{23} (Numsibs-> Ach)	-.060	-1.95	.050
B_{24} (VRQ->Ach)	.639	27.61	<.0001

Second Run of the Scottish Data - Inclusion of Second-Stage Predictors

A number of exploratory runs were made to determine which of the school level variables predict each path. As was mentioned earlier, since this is a full information maximum likelihood procedure and since predictors are multicollinear, inclusion or exclusion of a single predictor alters the entire solution. It is therefore necessary run numerous trials, testing whole sets of second-stage predictors. The end product of this exploratory phase is a quite modest model which has the form:

$$B_{11} = \bar{B}_{11} + \gamma_{111}(\text{Ave-VRQ}) + U_{11}$$
$$B_{12} = \bar{B}_{12} + U_{12}$$
$$B_{13} = \bar{B}_{13} + U_{13}$$
$$B_{21} = \bar{B}_{21} + U_{21}$$
$$B_{22} = \bar{B}_{22} + U_{22}$$
$$B_{23} = \bar{B}_{23} + \gamma_{231}(\text{Ave-SES}) + \gamma_{232}(\text{Ave-Occfath}) + U_{23}$$
$$B_{24} = \bar{B}_{24} + U_{24}$$

Only two paths have predictors. Table 3-A, for the baseline model, indicated that B_{12}, B_{13} and B_{24} had virtually no parameter variance and so were not susceptible to prediction. Two other paths (B_{21} and B_{22}) although having significant parameter variance, evidenced no relationship with the available set of school level predictors. This raises the possibility that important school level processes

are not represented and that the model may be misspecified at the second stage.

Another issue is purely statistical. The degrees of freedom are quite small in relationship to the number of parameters being estimated. There are ten fixed effects and only twenty schools. With more groups a more predictive between-group model may have been possible.

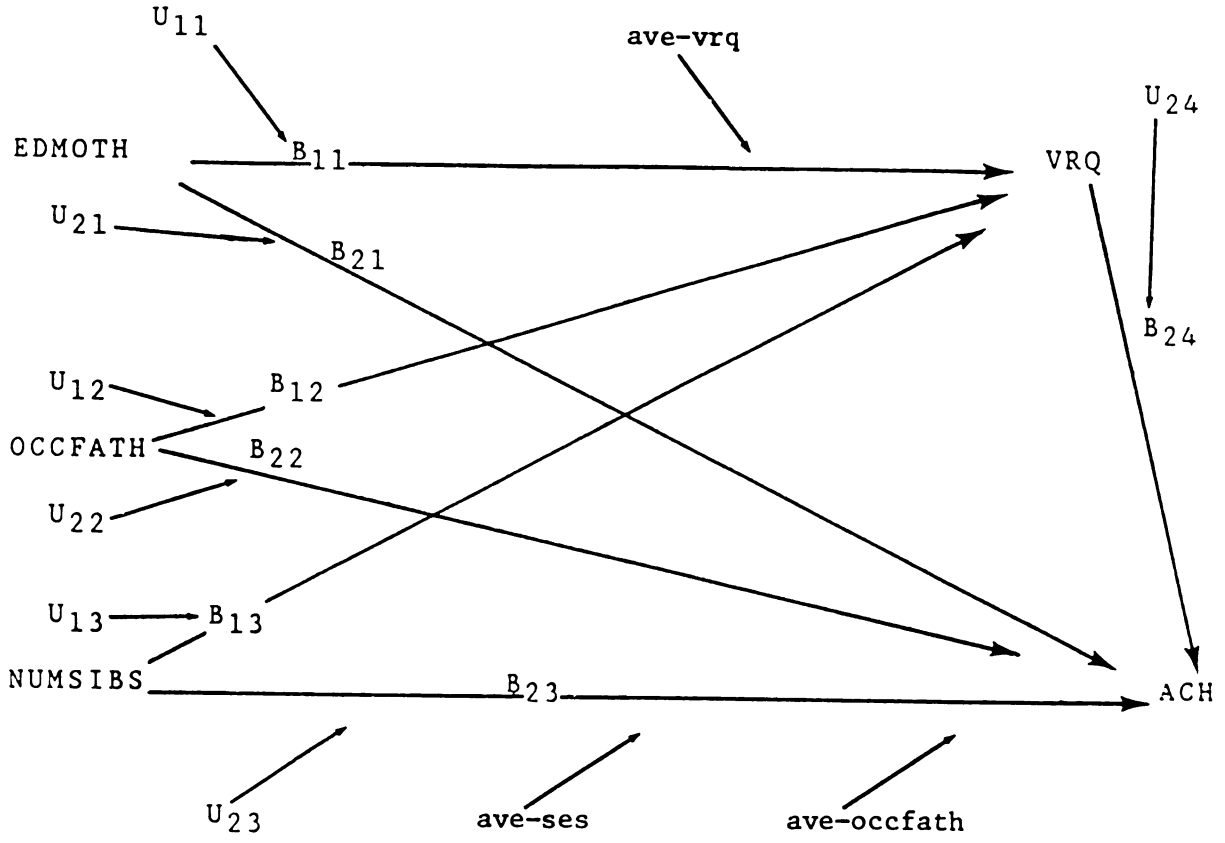


Figure 4.7
Scottish School Data: Structured Model

The B_{11} (Edmoth-->VRQ) path is explained by average school VRQ. The second-stage regression coefficient (table 4-C) is .25, implying that in schools with a high average VRQ, mother's education is more predictive of a student's verbal reasoning than in schools with a low average VRQ. Why this is the case is a matter of speculation. Perhaps average school VRQ is indicative of a facilitative school learning atmosphere where a mother's contribution to the general education of her children will be reinforced rather than drowned out. More in-depth information on the nature of the school processes would be required to illuminate this question. Whatever the underlying mechanism, Ave-VRQ explained 46 percent of the total parameter variance, as is indicated by table 4-A. Also, the chi-square probability of the conditional parameter variance is .06, which is non-significant by a strict criterion. So the B_{11} path has been substantially explained by the model.

A rather different result was found in the second-stage prediction model for B_{23} (Numsibs-->Ach). Table 4-C shows that average school SES has a second-stage regression coefficient of .61 for predicting B_{23}. Since the average B_{23} path is negative (-.06) higher school SES would tend to make this path less negative, or more positive. School SES has a large enough standard deviation that in the highest SES schools the B_{23} path would flip around and become positive. Perhaps this is a result of a threshold effect. If the

family is financially well off having siblings increases opportunities for a child to learn. But below a certain economic threshold, a bigger family means greater financial burden and greater deprivation for the student. Again, only process information about schools can begin to answer these questions.

Another school level variable predicts the B_{23} (Numsibs-->Ach) path, namely the average occupational status of father. Surprisingly, this has a negative predictive coefficient for B_{23} which equals -.04. The Z test for this coefficient is significant at the .01 level. Why an SES-related variable would predict with an opposite sign as average SES is mysterious. This suggests that the model is not fully specified. If other relevant school level variables could have been added to the model, such an anomaly might disappear. Table 4-A indicates that the conditional parameter variance of this path is significant So, although the present second-stage model accounted for 38 percent of the parameter variance, there is more that can be explained.

In sum, table 4-A shows us that three paths (B_{12}, B_{13} and B_{24}) had virtually no between-group (parameter) variation. These paths can be regarded as constant over schools. One path, B_{11}, had virtually all of its between-group variation explained. Another path, B_{23}, had only part of its parameter variance accounted for. Two paths, B_{21} and B_{22}, had a significant amount of parameter variance but none of it was explained in a second-stage model.

Table 4-A

Scottish School Data

Parameter Variance of Paths

Structured Between-Group Model

Path	Parameter Variance	Chi-Square Statistic	Probability	Parameter To Total Variance	R^2
B_{11}	.01023	29.43	.06	.43	.46
B_{12}	.00401	16.81	.67	.31	——
B_{13}	.00272	13.86	.84	.27	——
B_{21}	.00580	33.48	.03	.42	——
B_{22}	.00743	35.40	.02	.49	——
B_{23}	.00683	29.96	.04	.52	.38
B_{24}	.00287	22.16	.33	.31	——

Table 4-B

Scottish School Data

Average Value of Paths

Structured Between-Group Model

Path	Average Value	Z Statistic	Probability
B_{11} (Edmoth-> VRQ)	.087	2.48	.013
B_{12} (Occfath-> VRQ)	.232	7.59	<.0001
B_{13} (Numsibs-> VRQ)	-.168	-5.85	<.0001
B_{21} (Edmoth-> Ach)	.095	3.61	.0003
B_{22} (Occfath-> Ach)	.112	3.99	<.0001
B_{23} (Numsibs-> Ach)	-.059	-2.06	.027
B_{24} (VRQ->Ach)	.639	27.02	<.0001

Table 4-C

Scottish School Data

Second-Stage Regression Coefficients

Path Predicted	Second Stage Predictor	Second Stage Regression Coefficient	Z Statistic	Probability
B_{11}				
	Ave-VRQ	.025	4.21	<.0001
B_{12}	———			
B_{13}	———			
B_{21}	———			
B_{22}	———			
B_{23}				
	Ave-SES	.607	3.36	.0008
	Ave-Occfath	-.040	-2.48	.013
B_{24}	———			

Although the inability of the analysis to explain much of the between-group variation in paths indicates an incomplete model, a baseline model is valuable in a multilevel path analysis. If our interest lies getting precise estimates of a path model for each group, a multilevel path analysis yields posterior estimates of paths which have the smallest possible means square error. If our interest lies in explaining paths rather than estimating them, a rich and correctly specified second-stage model is a necessity.

The analyses of both datasets has illustrated the usefulness in stipulating path models at the between-group level. In both cases the path model made substantive sense and yielded sensible results after the analysis was performed. The multilevel path analysis was less successful in explaining the between-group variance of the paths. This difficulty is symptomatic of the fact that the illustrations offered here were borrowed from datasets that were designed for other purposes. If the possibilities of multilevel path analysis are going to be fully realized in the future, studies will have to be designed for the purpose of explicating a path model in numerous groups. This means that a rich mix of process related variables has to be gathered at all levels of analysis. When there is a more thorough matching of statistical modeling and research design, substantive theory will be better informed.

CHAPTER V:

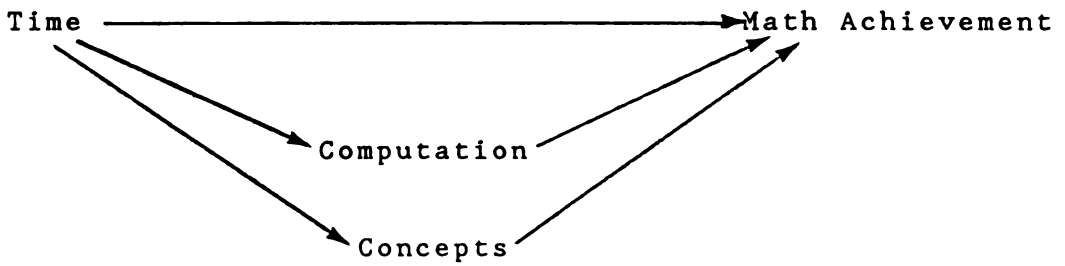
CONCLUSION

I. Introduction

In the preceding chapters it has been argued that a multilevel path model would merge the statistical traditions of multilevel analysis and path analysis into a single powerful analytic tool. In chapter two a statistical model was derived which represented one type of multilevel path model. Chapter three reviewed issues associated with the production of a computer algorithm which would create estimates derived from the model. Finally analyses of actual datasets were presented in chapter four utilizing a computer program written according to the principals outlined in chapter three. The analysis section demonstrated that the multilevel path model gives interpretable results when applied to the sorts of datasets that occur in large-scale educational studies.

In principle, the multilevel path approach would be pertinent whenever information is presented to the researcher at two levels of analysis and the researcher is interested in deducing causal processes at the 'lower' level. The most obvious example of this is a study of students nested within numerous schools, the situation in both datasets analyzed in chapter four. In this instance we are interested in modelling processes within each of numerous schools. A less obvious example would be to study a set of individuals observed at numerous time points. In the previous example a

group was represented by observations on numerous individuals within a school. In this example the 'group' is represented by observations at numerous time points within an individual. If we were studying what contributes to the development of math skill in early elementary students, we might collect data on computational skills, understanding of concepts and general math ability at ten time points. A within-student path model might take the following form:



From such a model we could determine which, if any, math skills are important for the development of math ability. This application of the multilevel path model parallels the application of the Hierarchical Linear Model to individual growth curves, as outlined by Bryk and Raudenbush (1987).

These examples suggest that there is a wide range of applicability for the multilevel path models. Nevertheless, such models are especially useful if certain conditions are met:

1. There are large datasets. Experience with the related HLM has shown that data from tens of groups is required in order to give precise estimates (Raudenbush, 1984). In

fact, for a given total sample size, it is better to have many small groups than a few large groups.

2. Similar processes occur in all groups. It is assumed in the multilevel path model that the same variables are related by the same causal network in all groups. In a sense, each set of within-group paths represents a replication of path model experiment.

3. If there is information available about the nature of the groups, this information must explain why processes are different from one group to another. This is because such variables serve as predictors in a between-group model which models variation in the within-group paths.

II. Problems With the Multilevel Path Model

Practicality

Although the results in chapter four demonstrated that application of the multilevel path model can lead to interesting results, this methodology may not always be feasible because the algorithm is computationally intensive. The EM algorithm requires many passes through the data before it converges, so without a fast mainframe computer and a healthy computer budget, such techniques may be impractical. Ironically, one way around this problem may be the 'low tech' approach. If a version of the estimating program could be written to work on a micro computer, expense would not be a factor. The analysis could be set into motion and

the computer could be left on its own for however long is required. This could be cumbersome from a time standpoint though.

Increased Burden of Proper Specification

Another problem with this type of analysis is a conceptual, rather than a practical, one. Multilevel path models will give proper estimates only if the models are properly specified at both levels of analysis. This imposes a heavy a priori burden on theory and emphasizes that this technique is not particularly appropriate for exploratory purposes.

Statistical Tests

A third problem that comes to light is statistical in nature. The chi-square test for parameter variances is only approximate. One reason for this is because the statistic is a function of an estimated regression coefficient that is the least squares estimate of paths and involves the inversion of the data matrix for each equation, for each group. There will often be groups that do not have full rank predictor matrices. For example, if gender is a predictor and a group is composed of all females, the variance and covariance terms for gender will be zero. At present, non-full-rank cases are simply excluded from the calculation of the statistic, but not from the Bayesian estimates (which do not require the inversion of data matrices). So the estimates of

parameters and the test statistics for parameter variances may be based on two sets of groups.

Another reason that test statistics are approximate is because they treat dispersion parameters as known quantities. The error of the dispersion estimates cannot be estimated by the present program.

III. Limitations

Limited Model Definition

The main limitation of this version of the multilevel path model is that it represents only one of many feasible configurations. Muthen and Satorra (1987) define the conceptual possibilities for defining multilevel structural models which include: 1) measurement models at the first and/or second stage 2) random predictors at the first and/or second stage 3) A path model at the second stage. These possibilities define 32 different configurations which would be possible for multilevel path model, of which the present model is one.

Lack of a Test of Fit

In any path analysis it is very useful to have some criterion by which to assess how well the model fits the data. In a LISREL model one can test the fit of the model by employing an omnibus chi-square test based on a likelihood

ratio test (Joreskog, 1973). An analogous test has not been developed for the multilevel path analysis. Perhaps a likelihood ratio test based on the log likelihood of the data (derived in chapter two) could be devised.

Uncorrelated Disturbances

The next issue may not be a limitation as such, but a strong assumption. In the model devised in this thesis, the first-stage disturbances (or errors) are uncorrelated, meaning that ψ is diagonal. This means that the R_k are uncorrelated for the within-group equation system (individual i and group j);

$$Y = Z\,B + R =$$

$$\text{equation 1} \quad Y_1 = X_1\, b_{(x)11} + X_2\, b_{(x)12} + \ldots + X_q\, b_{(x)1q} + R_1$$

$$\text{equation 2} \quad Y_2 = Y_1\, b_{(y)21} + X_1\, b_{(x)21} + \ldots + X_q\, b_{(x)2q} + R_2$$

$$\vdots \qquad\qquad \vdots$$

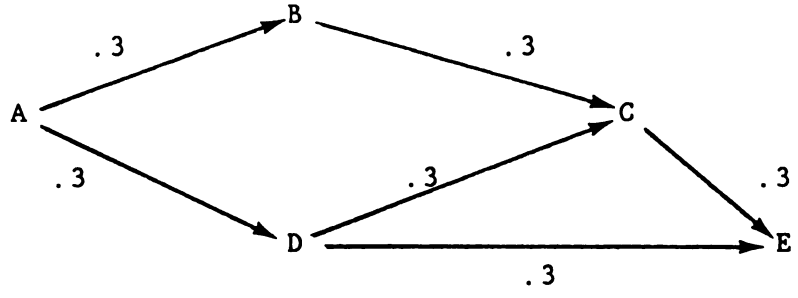
$$\text{equation K} \quad Y_K = Y_1\, b_{(y)K1} + Y_2\, b_{(y)K2} + \ldots + Y_{(K-1)}\, b_{(y)K(K-1)} + X_1\, b_{(x)K1} + \ldots + X_q\, b_{(x)Kq} + R_K$$

This is a fortuitous assumption because it enables us to use the separate-equation regression approach to estimate the paths (Land, 1973). The assumption of uncorrelated

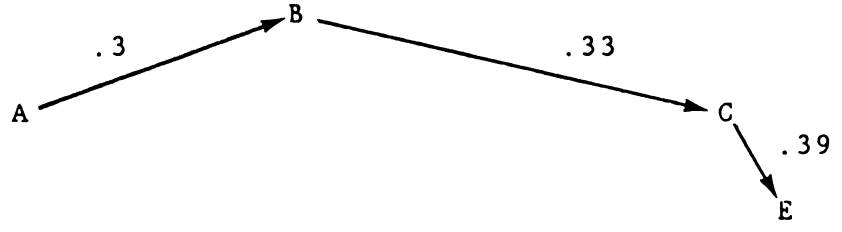
disturbances is perfectly reasonable assuming that all pertinent variables are in the model and the configuration of paths is correct. It is commonly believed that if there is a variable missing from the path model and that this variable has a causal influence on two or more endogenous (outcome) variables, the disturbances will be correlated. This sentiment is echoed by Hanushek and Jackson (1977) "If the same explanatory factor is excluded from more than one equation, the effect of that factor will be present in more than one error term and will cause the error terms to be somewhat correlated" (p. 230). In Joreskog's LISREL model (Joreskog & Sorbom, 1978) one can allow the error terms, R_k, to be correlated. The motivation behind doing so is to make up for such missing, confounding predictors (confounding in that the missing variable is related to at least two endogenous variables). It is assumed that since <u>some</u> variables are almost always left out of any model, correlated error terms are a way to represent the effect of these missing variables and the result will be a model that fits the data well.

Hunter and Gerbing have disputed this claim and have devised two counter examples to disprove it (Hunter & Gerbing, 1981). The first counterexample illustrates a situation in which a confounding variable is left out of the model but the resultant path model has uncorrelated errors. The model can remain well specified even in the face of missing confounding variables and uncorrelated disturbances if paths are added to the model which make the connections which

would have been made if the missing variables had been in the model. These connections will be indirect causal paths because they would have been mediated by a missing variable. As an example I will give a simplified version of Hunter and Gerbing's illustration. Suppose the complete causal system pictured below. The values are the causal paths for the population:

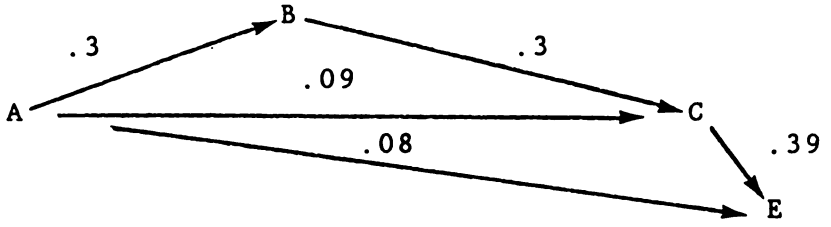


Now suppose the path model that is specified leaves out factor 'D'. The usual custom is to leave out all direct paths associated with 'D'. This leads to the following estimate:

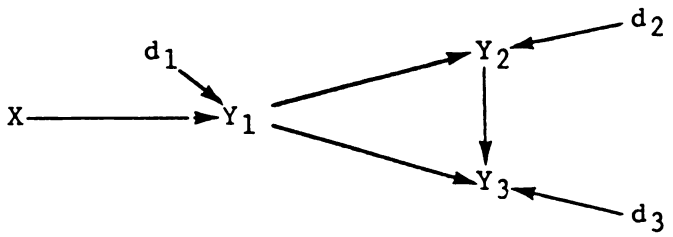


This is a poor fitting model but it could be 'fixed up' by allowing the disturbances to be correlated. But another

model could have been specified which would not necessitate correlated disturbances by simply putting the indirect paths which would have been mediated by 'D':

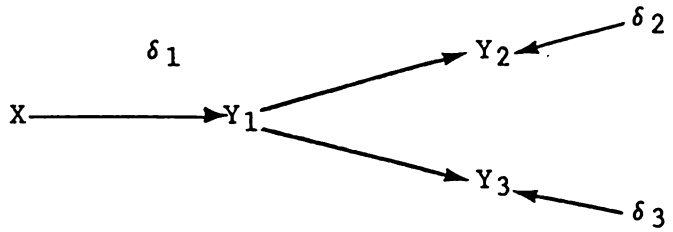


One often expects there to be missing intervening variables, but these can be accommodated by the correct specification of paths. In a second counterexample Hunter and Gerbing illustrate a situation in which a misspecified model is made to display apparent good fit by incorrectly allowing disturbances of the equations to be correlated. First we have the actual model;



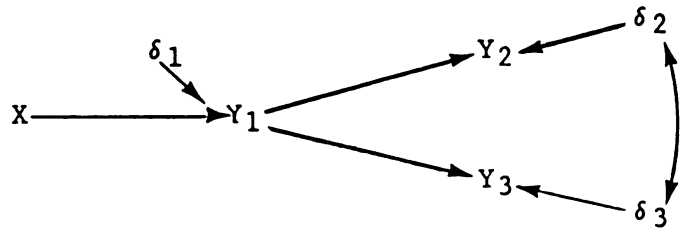
Where d_1, d_2 and d_3 are uncorrelated.

A misspecified model will be defined if we leave out the path from Y_2 to Y_3;



Hunter and Gerbing claim that a LISREL analysis on such a misspecified model will not fit the data well in the sense that it will not reproduce observed correlation matrix.

But what if we further misspecify the model by stipulating that the disturbance terms for Y_2 and Y_3 are correlated (as is indicated by a curved arrow);



In this case they claim that a LISREL analysis fits the data almost perfectly. The moral is that specifying correlated disturbances does not make up for missing variables. It may instead cover up misspecified paths.

The upshot of these examples is that there is no analytic substitute for a properly specified path model; one that has properly defined paths as well as variables. As a result, in multilevel path models even more onus is placed on proper model specification.

In the less ideal world of actual data analysis the equations may in fact be correlated to some extent. It would be quite useful in the future to do monte carlo studies to determine how results are effected by mild departures from the assumption of uncorrelated errors.

IV. Future Work

Defining Fixed Effects

The multilevel path model defined in this thesis is an instance of a mixed model, i.e. both fixed and random effects are present in the combined model (Braun, Rubin and Thayer, 1983). From chapter two the mixed model is given as:

$$Y = A_1\theta_1 + A_2\theta_2 + R \ , \qquad (5.1)$$

where γ is the fixed effect and U is the random effect. The mixed model is expressed in the combined-level form by,

$$Y = ZW\gamma + ZU + R \ . \qquad (5.2)$$

This in turn can be separated into its two-stage hierarchical form by equations:

$$Y = ZB + R \ , \qquad (5.3)$$

$$B = W\gamma + U \ . \qquad (5.4)$$

By substituting 5.4 into 5.3 we get Equation 5.2. In the combined form it is apparent that the vector U ,containing the

random effects, constitutes the parameter differences across groups and the vector γ, containing the fixed effects, constitutes the second-stage regression coefficients. If the chi-square test indicates that one of the parameter variances is zero, this implies that the parameter error, U_{pj}, corresponding to path B_{pj}, is zero in every group j. In other words, if a path has zero parameter variance, the random component constituting the path is null so the second-stage model for path p is simply:

$$B_p = W\gamma$$

The advantage of the mixed-model formulation is that a first-stage parameter can be modeled as a fixed effect by eliminating the corresponding U_{pj} in every group. If J is the number of groups this can be accomplished by deleting the J U_{pj} elements from U, and suitably reducing the column dimension of A_2 (or Z) by J. When some paths in fact have no variance, estimates assuming that these paths are fixed are more valid than estimates assuming that these paths are random. The programming needed to fix effects will be the next feature added to the multilevel mixed model. The analyses presented in this thesis will be redone with the appropriate paths defined as fixed effects.

Adding Intercepts to the Within-Group Path Model

Another elaboration that can readily be added to multilevel path analysis is the addition of intercepts to

the first-stage model. If predictors are also mean deviated, the intercepts will be group means which are interesting from a substantive point of view. Presently, the multilevel path program has the option to add group-mean intercepts to the path model. This elaboration wasn't presented in the current analysis because paths were the main focus of the thesis. The path model for the High School and Beyond defined for intercepts would have the form (for individual i and group j):

$$\text{Classes} = \overline{\text{Classes}} + B_{11}(\text{Minority} - \overline{\text{Minority}}) + B_{12}(\text{SES} - \text{SES}) + R_1$$

$$\text{Achievement} = \overline{\text{Achievement}} + B_{21}(\text{Classes} - \overline{\text{Classes}}) + B_{22}(\text{Minority} - \overline{\text{Minority}}) + B_{23}(\text{SES} - \overline{\text{SES}}) + R_2$$

The advantage of adding group means is that they can be analyzed in a second-stage model so that we can define the antecedents of group effects as well as group processes. For example, it would be interesting to know if Catholic schools were more equitable (smaller SES->Ach path) but at the expense of average achievement for the school. The data analysis presented in this thesis will be rerun with group means in the near future.

Inclusion of a Measurement Model

One of the possibilities summarized by Muthen and Satorra (1987) involved adding a measurement model to the within-group model. Ignoring measurement error at the first stage tends to bias estimates of paths and inflate the estimates of sampling error. The notion of a measurement model in a multilevel context reiterates many of the issues that arose with the concept of a within-group path modelling. Do we assume that the factor structure is the same for all groups? Do we further assume that the factor loadings constant across groups? Parallel to the path model formulation, it is my judgement that the factor structure will be constant across groups but the actual loadings will vary. This implies that a measurement model would be formulated by running separate confirmatory factor analyses which test the same factor structure, in all groups.

Another central question is how would one best enact a measurement model simultaneously defined over numerous groups. The LISREL program currently has the capability of estimating a measurement model. But executing a separate LISREL analysis in each of over a hundred groups would be computationally prohibitive. Also, LISREL is based on large sample estimating theory which might be inappropriate for the often small group size to be found in multi-group data bases.

There is another, more conceptual, objection to a LISREL type approach. LISREL simultaneously estimates measurement

coefficients and path coefficients. Since LISREL is based on full information maximum likelihood (as is the multilevel path program) misspecification errors will tend to bias the path estimates and visa versa. Heise says of full information maximum likelihood (FIML) estimation "All the observed variances and covariances simultaneously contribute to the estimation of all the parameters...FIML methods are quite sensitive to specification error" (Heise, 1975). This is an especially acute problem if the definition of the measurement model is in an exploratory phase where different factor structures are being piloted or items are being assessed for feasibility of inclusion in scales. Imbedding the measurement part of the model within a larger two-stage path model would mean that an independent assessment of the quality of the measurement instruments could not be had.

Gerbing and Hunter site an example where a deliberately misspecified path model gives incorrect estimates of factor correlations, "Even though the error was in the causal mode, LISREL placed all of the error into the estimated factor correlations and maintained perfect consistency between factor correlations and the incorrect path coefficients" (Gerbing & Hunter, 1980). They conclude that the simultaneous analysis of measurement and causal models may be suited for correctly specified models but "There is no a priori reason why a researcher would induce the additional complexity of simultaneous analysis of untested measurement and causal

models except that the necessary machinery exits for such an analysis" (p19).

This problem can be avoided if measurement model definition is a wholly separate phase of the analysis. This gives the researcher an opportunity to troubleshoot scales through numerous confirmatory runs. When a valid measurement structure is confirmed, the latent covariances of factors are input into the path analysis. This approach was successfully employed by the author (Jenkins, 1985) in defining a large scale measurement model for later input into LISREL. The approach in that study was to first define the measurement structure through a least squares confirmatory factor analysis procedure. This structure was then validated by a confirmatory factor analysis using LISREL. Interestingly the least squares and LISREL runs gave similar factor loadings.

A measurement model as a separate stage of analysis would be relatively straightforward to implement with the present estimation program. A separate routine could be written to implement a least square confirmatory factor analysis in all groups simultaneously (see Hunter & Gerbing, 1979). Some summary statistics to represent fit of the model would have to be devised (e.g. average residual correlations; means, maximum and minimum values of loadings, average factor/factor correlations, etc.). Variables will be added to and deleted till a good-fitting factor structure has been defined. Then the estimated factor/factor correlations for each group would be fed into the multilevel path program as it exists now.

Between-Group Measurement Model

The issues involved with defining a between-group measurement model are simpler because only one model has to be devised, rather than one for every group. If a separate measurement model analysis were planned, it could be done via LISREL or a least squares confirmatory package (Hunter & Gerbing, 1979) and the resulting factor scores could be fed into the existing multilevel path program.

An attempt to simultaneously define a measurement and a path model would be part of an effort to define a group-level path model. This will be discussed in the next section.

Group-Level Path Model

It is not immediately apparent that defining a path model at the group level would be useful. For a group-level path model first-stage paths would be the second-stage endogenous variables. I cannot think of a sensible interpretation for a situation where one within-group path causes another such path. A between-group path model would be sensible under two conditions a) we want to represent a network of relationships among the group-level variables and b) intercepts (which are group means) are included in the model as predictors in the second-stage path analysis. These sorts of intercepts are the same as group-level variables.

The easiest way to devise a between-group path model would be to repeat the same general approach used for the within-group path model. This would involve defining a simultaneous equation system in which paths were outcomes and group level variables (or intercepts) are predictors or outcomes. As with the first-level model, the errors of the equations would be uncorrelated

For example consider again the first-stage path model for the High School and Beyond data. This model, defined with intercepts and assuming mean-deviated predictors takes the following form for individual i and group j:

$$\text{Classes} = \overline{\text{Classes}} + B_{11}(\text{Minority}) + B_{12}(\text{SES}) + R_1$$

$$\text{Ach} = \overline{\text{Ach}} + B_{21}(\text{Classes}) + B_{22}(\text{minority}) + B_{23}(\text{SES}) + R_2$$

There are four variables defined at the group level; Sector, Ave-SES, Sd-SES, and Ave-Classes. Note that Ave-Classes and the Classes intercept would be the same variable. A Possible second-stage path model might be defined as follows:

$$\text{Sector} = \gamma_{10}(\text{Ave-SES}) + \gamma_{12}(\overline{\text{Classes}}) + \gamma_{13}(\text{Sd-SES}) + U_{11}$$

$$B_{11} = \gamma_{20} + \gamma_{21}(\text{Sector}) + U_{12}$$

$$B_{12} = \gamma_{30} + \gamma_{31}(\text{Sector}) + \gamma_{32}(\text{Ave-SES}) + U_{12}$$

$$B_{21} = \gamma_{40} + \gamma_{41}(\text{Sector}) + U_{21}$$

$$B_{22} = \gamma_{50} + \gamma_{51}(\text{Sd-SES}) + U_{22}$$

$$B_{23} = \gamma_{60} + U_{23}$$

Here we see that the model previously defined for the paths is as it was before, but now there are relationships

among the group-level variables and intercepts. Specifically, all the group-level variables which characterized Sector are explicitly introduced as exogenous predictors (the first equation).

The complexity has increased quite a bit over the previously defined multilevel path model. Possibly the greatest problem with these models will be to keep them simple enough.

This sort of scheme would fit into a second-stage model which, in matrix terms, looks the same as before:

$$B^* = W\gamma + U$$

But unlike the previous formulation the outcome vector, B^*, contains more than just paths, it can contain paths, intercepts and group-level variables. As before for group j, $Var(U_j) = T$. But now T is defined as a diagonal matrix, (i.e. errors are not correlated) for the same reason that ψ is diagonal; disturbances are independent in a properly specified path system. This simple scheme for a group-level path model would fit into the present estimating program with little renovation.

Full Blown Path Analysis

One may ask, "Why not just do a LISREL model at the group level?" This is a theoretical possibility, but a between-group LISREL model would depart from the statistical assumptions

of the present estimation theory. First of all in the LISREL model it is assumed that exogenous variables are random. In the General Bayesian Linear Model exogenous variables are fixed. Also, with LISREL path coefficients are defined in two rectangular matrices of fixed parameters. In contrast, with the Bayesian approach second-stage path coefficients are defined as a vector of random coefficients with a vague prior distribution. At present the LISREL model would not fit into the context of the General Bayesian Linear Model. The assumption of random exogenous variables at the first or second stage of the hierarchy would necessitate a reformulation of the multilevel path model, possibly along lines other than those developed in chapter two.

This work represents a beginning of the actual estimation of multilevel path models. This final chapter has suggested a few ways the present statistical approach could be extended. There are doubtless other approaches which would further expand the scope of these models.

Regardless of the analytic form such approaches take in the future, I hope to have demonstrated that multilevel path models have promise to be a powerful tool for illuminating important issues in social science research.

APPENDIX

APPENDIX

DERIVING THE POSTERIOR DENSITY FUNCTION OF θ

From Equation 2.14 the posterior density of θ is:

$$f(\theta|Y,\Psi,\Omega,A) =$$
$$(2\pi)^{-KN/2}\ (2\pi)^{-t/2}\ |\Psi|^{-1/2}\ |\Omega|^{-1/2}$$
$$\exp\{-1/2(Y-A\theta)'\ \Psi^{-1}\ (Y-A\theta)\}\ \exp\{-1/2\theta'\Omega^{-1}\theta\} \quad (A.1)$$

The exponential componant is the quadratic Q :

$$Q = (Y-A\theta)'\Psi^{-1}(Y-A\theta) + \theta'\Omega^{-1}\theta \quad (A.2)$$

Expanding terms we get:

$$Q = Y'\Psi^{-1}Y - Y'\Psi^{-1}A\theta - \theta'A'\Psi^{-1}Y + \theta'A'\Psi^{-1}A\theta + \theta'\Omega^{-1}\theta \quad (A.3)$$

The first term, $Y'\Psi^{-1}Y$,is not a function of θ and so is a constant with respect to the density. The corresponding term, $\exp\{-1/2Y'\Psi^{-1}Y\}$, is taken out of the exponent and put into the constant term. The remaining terms are combined and arranged in descending powers of θ to yield:

$$Q = \theta'(A'\Psi^{-1}A+\Omega^{-1})\theta - 2Y'\Psi^{-1}A\theta \quad (A.4)$$

Q now has the general quadradic form:

$$Q = X'MX - 2B'X \quad , \quad \text{with :} \tag{A.5}$$

1) $X = \theta$

2) $M = (A'\Psi^{-1}A+\Omega^{-1})$

3) $B = A'\Omega^{-1}Y$

The square of the quadratic can be completed to put Q into the algebraically equivalent form:

$$Q = (X-MB)'M(X-MB) - B'MB \tag{A.6}$$

Substituting for X, M and B we get:

$$\begin{aligned} Q = {} & [\theta - (A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y]'(A'\Psi^{-1}A+\Omega^{-1}) \\ & [\theta - (A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y] - Y'\Psi^{-1}A(A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y \end{aligned} \tag{A.7}$$

The last term, not being a function of θ, can be taken out of the quadratic and put into the constant to yield:

$$\begin{aligned} Q = {} & [\theta - (A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y]'(A'\Psi^{-1}A+\Omega^{-1}) \\ & [\theta - (A'\Psi^{-1}A+\Omega^{-1})A'\Psi^{-1}Y]) \end{aligned} \tag{A.8}$$

Which is the result required for Equation 2.15 in the text.

BIBLIOGRAPHY

BIBLIOGRAPHY

Barcikowski, R. (1981). Statistical power with group mean. Journal of Educational Statistics, 6(3), 267-285.

Bianchi, L.(1987). Estimating the covariance componants of an unbalanced multivariate latent random model via the EM algorithm. Unpublished dissertation, College of Education, Michigan State University.

Bishop, Fineburg & Holland (1975). The D method for calculating asymptotic distributions. In Chpt. 14.6, Discrete Multivariate Analysis. Cambridge, Mass, MIT Press.

Braun, H., Jones, D., & Rubin, D. (1982). Empirical Bayes estimation of coefficients in the general linear model with data of deficient rank. Psychometrika 48(2), 171-181.

Bryk, A. & Raudenbush, S. (1987). Application of hierarchical linear models to assessing change. Psychological Bulletin 101(1), 147-158.

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

Burstein, L., Linn, R. & Capell, F.(1978). Analyzing multilevel data in the presence of heterogenous within-class regressions. Journal of Educational Statistics, 3(4),347-383.

Bryk, A., Raudenbush, S., Seltzer, M. & Congdon, R.(1986). An Introduction to HLM: Computer Program and User's Guide. University of Chicago Dept of Education.

Campbell, D. & Stanley, J. (1966). Experimental and Quasi-experimental Designs for Research. Rand McNally college Publishing Co., Chicago.

Coleman, J., Hoffer, T. & Kilgore, S. (1982). High School Achievement: Public, Catholic, and Private Schools Compared. N.Y.: Basic Books.

Cooley, W., Bond, L. & Mao, B. (1981). Analyzing multi-level data. In Berk, R.A. (Ed) Educational Evaluation Methodology. Baltimore: Johns Hopkins University Press, 64-83.

Cronbach, L. & Webb, W. (1975). Between and within-class effects in a reported aptitude by-treatment interaction: reanalysis of A Study by G. L. Anderson. Journal of Educational Psychology, 6, 712-724.

Dempster, A., Laird, N. & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm Journal of the Royal Statistical Society, series B V. 34, 1-8.

Demster, A., Rubin, D. & Tsutakauwa, R. (1981). Estimation in covariant component models. Journal of the American Statistical Association, 76, 341-353.

Duncan, O., Featherman, D. (1973). Psychological and cultural barter in the Process of Occupational Structural Equations in the Social Science, Guildbergh & Duncan (Eds). Seminar Press, N. Y.

Efron, B. & Morris, C. (1977). Stein's paradox in statistics, Scientific American. 36(5), 119-127.

Gerbing, D. & Hunter, J. (1980). The return to multiple groups: an analysis and critique of confirmatory factor analysis with LISREL. Manuscript presented to the Southwestern Psychological Association.

Glass,.G. & Stanley, J.(1970). Statistical Methods In Education and Pshchology, Englewood Cliffs, N.J.:Prentice-Hall.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. Biometrika 73(1), 43-56.

Hanushek, E.(1974). Efficient estimators for regressing regression coefficients. The American Statistician, 28(2),66-67.

Hanushek, E. (1977). Statistical Methods for Social Scientists New York: Academic Press, Inc.

Hartley, H. O. (1958). Maximum likelihood estimation from incomplete data Biometrics (June).

Hartley, H. W., Hocking, R. R. (1971). The analysis of incomplete data Biometrics (Dec).

Harville, David (1977). Maximum likelihood approaches to variance component estimation and to related problems. Journal of the American Statistical Association 72, 358.

Hedges, Larry (1982). Estimation of effect size from a series of independent experiments. Psychological Bulletin 92, 490-499.

Heise, David (1975). Causal Analysis John Wiley & Sons, N. Y.

Henderson, Charles & Henderson C. (1979). Analysis of covariance in mixed models with unequal subclass numbers. Commun. Statist. Theory lMeth. A8(8), 751-787.

Hoel, P. ,Port, S. & Stone, C.(1971). Introduction to Probability Theory. Houghten Mifflin Co., Boston.

Hoffer, T. Greeley, A. & Coleman, J.(1985). Achievement growth in public and Catholic schools. Sociology of Education, 58, 74-97.

Hopkins, Kennith (1982). The unit of analysis: group means versus individual observations. American Educational Research Journal 19(1), 5-18.

Houang, R. & Schmidt, W. (1981). A comparison of three analytical strategies for hierarchical data. Revision of a paper presented at the annual meeting of the American Educational Research Association Meeting, Los Angeles.

Hui, S. & Berger, J. (1983). Empirical Bayes estimation of rates in longitudinal studies. Journal of the American Statistical Society (Dec).

Hunter, J. & Gerbing, D. (1979). Unideminsional measurement and confirmatory factor analysis, Occasional Paper: The Institute for Research on Teaching, Michigan State University.

Hunter, J. (1980). The dimensionality of the general aptitude test battery (GATE) and the dominance of general factors over specific factors. Draft Manuscript, Dept. of Psychology, Michigan State University.

Hunter, J. & Gerbing, D.(1980). Unidimensional measurement, second order factor analysis and causal models. In Stran & Cummings (Eds.). Research in Organizational Behavior, 4, Greenwhich Conn: Jai Press.

Jenkins, F. (1985). Defining a general classroom writing ability: a measurement model. Paper presented at the annual meeting of the American Educational Research Association, April 1985.

Jenkins, F. (1987). Path modeling with individual by group interactions. Paper presented at the annual meeting of the American Educational Research Association meeting, Washington, D.C.

Joreskog, K. (1973). A General method for estimating a linear structural equation system. In Goldberger & Duncan, (Eds), Structural Equation Models in the Social Sciences, Seminar Press, N. Y.

Joreskog, K. & Sorbom, D. (1978). Lisrel IV: Analysis of Linear Structural Equations by the Method of Maximum Likelihood. International Educational Services, Chicago.

Kasim, R. & Raudenbush, S. (1986). Examining variances in hierarchical models. Paper presented at the annual meeting of the American Educational Research Association Meeting, San Francisco, March 1986.

Knapp, J. (1977). The unit-of-analysis problem in applications of simple correlational analysis to educational research. Journal of Educational Statistics 2(3), 171-186.

Laird, N. & Ware, J.(1982). Random-effects models for longitudinal data. Biometrics, 38, 963-974.

Land, K. (1973). Identification, parameter estimation, and hypothesis testing in recursive sociological models. In Goldberger & Duncan, (Eds), Structural Equation Models in the Social Sciences, Seminar Press, N. Y.

Lee, V. (1986). Multi-level causal models for social class and achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lee, V. & Bryk, A.(1986). The effects of high school academic organization on the social distribution of achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Lindley, D. & Smith, A.(1972). Bayes estimation for the linear model. Journal of the Royal Statistical Society, Series B, 34,1-41.

Lindley, D. & Smith, A. (1982). Bayes estimates from the linear model. Journal of the Royal Stastical Society. 3(13), 1-41.

Longford, N. T. (1985). A fast survey algorithims for maximum likelihood estimation in unbalenced mixed models with retest effects. Unpublished manuscript, Institute for Applied Statistics, Lancaster University, Lancaster, England.

Mason,.W., Wong, G. & Entwisle, B.(1984). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), Sociological Methodology 1983-1984, 72-103, San Francisco: Jossey-Bass.

McNemar, Q. (1940). Review of Linquist's Statistical Analysis In Educational Research. Psychological Bulletin. 2(3), 747.

Mikhail, W. M. (1975). A comparative monte carlo study of the properties of economic estimators. Journal of the American Stastical Association (March) 70(349).

Morris, C. (1983). Parametric empirical Bayes inference, theory and applications Journal of the American Statistical Association 78, 47-65.

Morris, C. (1983). Parametric emprircal Bayes inference: theory and applications. Journal of the American Statistical Association. 78(381).

Morrison, D. (1976). Multivariate Statistical Methods. McGraw-Hill, N. Y.

Muthen, B. & Satorra, A. (1987). Multilevel aspects of varying parameters in structural models. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C.

Page, E. B.(1975). Statistically recapturing the richness within the classroom. Psychology in the Schools, 12, 339-344.

Pellemer, D. & Light, R. (1980). Synthesizing outcomes: how to use research evidence from many studies. Harvard Educational Review (May) 50(2).

Raghu, D. & Harville, D. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. JASA (Dec).

Raudenbush, S. (1984-A). Application of a hierarchical linear model in educational research. Unpublished Doctoral Dissertation, Harvard University.

Raudenbush, S. (1984-B). Magnitude of teacher expectancy effects on pupil IQ as a function of the credibility of expectancy induction: a synthesis of findings Journal of Educational Psychology, 76(1), 85-97.

Raudenbush, S. (1988). Estimating change in dispersion. Journal of Educational Statistics, 13(2), 148-172.

Raudenbush, S. (1988). Educational applications of hierarchical linear models. Journal of Educational Statistics, 13(2), 85-116.

Raudenbush, S. (1987-A). Examining correlates of diversity Journal of Educational Statistics 12(3), 241-269.

Raudenbush, S. (1987-B). Likelihood formula for two and trhee level models, Unpublished Manuscript

Raudenbush, S. & Bryk, A. (1984). Application of emprircal Bayes estimation in educational research. Paper presented at the annual meeting of the AMerican Educational Research Association, New Orleans, 1984.

Raudenbush, S. & Bryk, A. (1985). Empirical bayes meta-analysis Journal of Educational Statistics 10(2), 75-98.

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

Raudenbush, S. & Bryk, A. (1988). Methodological advances in studying effects of schools and classrooms in student learning, Draft Manuscript to appear in Review of Research in Education.

Rubin, D.(1980). Using empirical Bayes techniques in the Law School Validity studies. Journal of the American Statistical Association, 75, 801-827.

Rubin, D. (1981). Estimation in parallel randomized experiments Journal of Educational Statistics (Winter) 6(4), 377-400.

SAS Institute Inc. (1985). The Matrix Procedure: Language and Applications. SAS Institute Inc., Gary, NC.

Searle, S. (1971). Linear Models. Wiley, New York.

Schmidt, W. (1969). Covariance structure analysis of the multivariate random effects model. Unpublished Dissertation, University of Chicago.

Smith, A. (1973). A general Bayesian linear model. Journal of the Royal Statistical Society, Series B, 35, 61-75.

Strenio, J.(1981). Empirical Bayes estimation for a hierarchical linear model. Unpublished dissertation, Department of Statistics, Harvard University.

Strenio, J., Weisberg, H. & Bryk, A.(1983). Empirical Bayes estimation of individual growth curve parameters and their relationship to covariates. Biometrics, 39,71-76.

Theil, H. (1971). The 2sls estimation method. In Chpt 9, Theil's Principals of Econometrics .

Walsh, J. (1947). Comparing schools in their examination performance: policy questions and data requirements. American Educational Research Journal (in press).

Wisenbaker, J. & Schmidt, W.(1979). The structural analysis of hierarchical data. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.