This is to certify that the

dissertation entitled

INCORPORATING FACTOR ANALYSIS INTO HIERARCHICAL MODELS

presented by

Yasuo Miyazaki

has been accepted towards fulfillment
of the requirements for

PH.D. degree in Counseling, Educational Psychology & Special Education

Major professor

Date 7/6/00

INCORPORATING FACTOR ANALYSIS INTO HIERARCHICAL MODELS

By

Yasuo Miyazaki

A DISSERTATION

Submitted to

Michigan State University

in partial fulfillment of the requirement

for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational

Psychology and Special Education

2000

ABSTRACT

INCORPORATING FACTOR ANALYSIS INTO HIERARCHICAL MODELS

By

Yasuo Miyazaki

This dissertation incorporates a factor analysis model into a two-level hierarchical linear model (HLM). It provides the model layout in HLM format and derives the maximum likelihood estimators. A computational program to implement the theory is developed. Special attention is focused on the application of the model in order to create a bridge between the statistical model and applications in education and human development. That is, I describe in detail when and in what context the model might be useful and the kinds of research questions that can be addressed using this model, so that researchers who are interested in substantive issues rather than statistical issues can immediately apply the model to their research.

Real as well as artificial data sets are used to illustrate the theory, the application, and the interpretation of the results. In the last Chapter, extensions of this model and their settings are mentioned.

DEDICATION

This dissertation is dedicated to my parents, Yoshio and Sakiko Miyazaki, for their unconditional love and support.

To my sister, Michiyo Kurakata, who cared about her elder brother, even though she was busy raising her four children.

To my late elder brother, Tomoo Miyazaki, who didn't have much happiness in his life, having passed away without reaching age 40, while I was in the United States. I apologized to him in front of his tomb for my absense, for not having come back when he died. That was three years ago already. This time I will bring this disseertation back home and will show it to him. He might ask, "Did you have fun in the U.S.?".

ACKNOWLEDGMENTS

It took a longer time than I thought to complete this dissertation. It was a long and lonely journey. At midnight, I often dreamed of the scenery of my home town village in my childhood, such as the gentle shape of the round mountains and the sound of the rice paddies when the breezes blew across the leaves of rice. The dream was often the same and was so vivid even though my childhood was more than 30 years ago. During tough times, I often spoke the golden words to myself, "There is no night that doesn't open (No night is endless)," and "Enduring snow, plums bloom sharper scents in the spring."

Additionally, the words from my father always relaxed me and made me realize that blue collar spirit was my blood: "There have been no persons in my family who can make a living by their brains. However, there is a way for such a person to survive. Use your body." The words from my mother, "I will live and wait until you come back home," lifted my spirit for completion of the program.

During my journey, I fortunately had many accompanying me. I would like to extend my sincere gratitude to my dissertation director, Professor Stephen W. Raudenbush, my advisor, Professor Kenneth A. Frank, and dissertation committee members, Professor Betsy J. Becker, Professor Mark Reckase, and Professor James Stapleton, for their advice, insight and guidance throughout the process of writing this dissertation.

My special thanks go to Dr. Randall P. Fotiu for his help with computer programming, which caused me a lot of trouble.

Haiyan Zhang spent lots of her time to help and to support me. Thank you very much, Haiyan.

I cannot enumerate here all the people to whom I am indebted, who offered me many kinds of help, support, and encouragement. I'm sure, however, that without their help, I wouldn't have finished the dissertation. I want to make this fact my constant reminder for the rest of my life. "Do a little help. That makes you a slightly better man."

TABLE OF CONTENTS

LIST OF TABLES

Chapter 1. Introduction

1-1. Motivation of the Study

Suppose that researchers are interested in how influences of students' gender, race, SES, family structure, IQ, and pretest score on their achievements vary from school to school. This is a typical research question that motivated the development of the hierarchical linear model (HLM) (Bryk & Raudenbush, 1992), which is alternatively called the multilevel model (Goldstein, 1995), the random coefficient model (Longford, 1993); and, often in statistics and biostatistics literature, the mixed model. (See, for example, SAS manual (1996) for Proc Mixed). The HLM can be implemented by several software programs, such as HLM (Bryk, Raudenbush, & Congdon, 1996), MLWin (Goldstein et al., 1998), SAS Proc Mixed (SAS Institute, 1996), and MIXOR (Hedeker, D., & Gibbons, R. D., 1996).

After the software became available, a fair amount of educational research was devoted to this issue (See, for example, Aitkin, Anderson, & Hinde (1980), Aitkin & Longford (1986), De Leeuw & Kreft (1986), Goldstein (1986), Raudenbush & Bryk (1986), and Lee & Bryk (1989) among others). Often, however, we don't have enough data to support a complex model that has many parameters. Specifically, if we increase the number of randomly varying coefficients, we need a larger sample size per cluster.

In an meta-analysis settings, suppose that we have multiple effect size estimates from a collection of independent studies that test the same hypotheses, and we wish to synthesize these effect sizes. In this setting, it is sometimes as important to make an inference about variance components as about fixed effects because the variances indicate

how much the effects sizes vary from study to study (Raudenbush, Becker, & Kalaian (1988), Kalaian & Raudenbush (1996)). Though in theory we can model as many the random effect sizes as we want, the number of the independent studies is usually relatively small in meta analysis (about 50 - 100 at most), naturally limiting for the number of random effects that can be estimated from the data at hand.

A more classical and standard application of multivariate random effects is a psychometric analysis for multivariate outcomes. For example, if researchers want to know the reliabilities of a test that is supposed to measure multiple constructs, we may formulate a two-level hierarchical model using dummy variables at level-1 as predictors. Then we ask how scores from multiple domains or items vary among students. From the variance estimates, we can determine the test reliabilities for each domain.

In the growth model literature within the hierarchical modeling framework, we often model the individual growth trajectory by a polynomial at level-1. At level-2 the coefficients of the level-1 model, that is, the person-specific growth parameters, become multivariate outcomes. In such a growth study, it is often researchers' substantive interest to examine correlations among the growth parameters, for example, the correlation between initial status and rate of growth. If the growth function is complex, there will be a large number of random effects at level 2.

The mixture of multivariate outcomes and growth models also applies to the context where we have many random effects in level-2. In this case, correlations within individuals occur for two reasons; one is that multiple measures are available for the same

individual and the other is that a single measure is taken over time for the same individual.

In a statistical sense, the common theme in the above examples is that the researchers wish to study many random effects simultaneously. If we have a large enough sample to estimate many parameters in the level-2 variance-covariance matrix (we will call this the "tau matrix"), it is possible to estimate such a large model. However, even in such a case, researchers who analyze data using a two-level hierarchical linear model still sometimes encounter the problem of slow convergence or even worse, noconvergence, because the tau matrix is really singular.

In statistical literature, the problem of a singular tau estimate is known as the Heywood case (Heywood (1931)) or boundary solution problem (Catchpole & Morgan (1994)). This problem occurs when the estimates reside near/at the boundary of parameter space. The currently most common approach to remedy this problem is to get rid of some of the random terms at level-2 and to re-estimate the smaller number of parameters in the tau matrix. Sometimes this approach does not make sense since removing a random term not only removes the large correlation that we want to get rid of, but also removes the variances that we want to keep in the model when the variation of the corresponding term exists.

Actually, we can find many examples of the above cases where the tau-matrix has high correlations among random terms. For example, Bryk & Raudenbush (P.141-144, Chapter 6, 1992) reanalyzed Huttenlocher et al. (1991)'s children's vocabulary growth data. The quadratic growth trajectory was formulated and they found the high correlations

among random coefficients of the initial status, slope, and rate of acceleration (Table 6.5, P.144, Bryk & Raudenbush, 1992), which was $\begin{pmatrix} 1.000 & -0.982 & -0.895 \\ -0.982 & 1.000 & 0.842 \\ -0.895 & 0.842 & 1.000 \end{pmatrix}$. Later the intercept was dropped completely from the model, but the correlation between the slope and the rate of acceleration was still very high at 0.904.

As a biostatistics example, Longford (P.136, 1993) analyzed the weights of newborn rats nested within litters and found that the correlation between the mean rat weights adjusted by diet contrasts and litter size and the gender gap in weight among litters was almost -1.0 ($\hat{\tau} = \begin{pmatrix} 0.505 & -0.139 \\ -0.139 & 0.038 \end{pmatrix}$, and the correlation is $\hat{\rho}(u_{0j}, u_{1j}) = \frac{\hat{\tau}_{01}}{\sqrt{\hat{\tau}_{00}\hat{\tau}_{11}}} = \frac{-0.139}{\sqrt{(0.505)(0.038)}} \cong -1.0034$, though this estimate is not admissible[1]).

In psychometrics, we are often interested in measuring the true score of the student. Suppose there is a general aptitude test that consists of 4 domains such as Math, Science, Reading, and Social Studies and each domain consists of 10 testlets whose scores are given simply by adding up the correct responses for the items in the testlet. If we conceive the above situation as testlets nested within students and we know which domain each testlet is supposed to measure, then there are four true scores for each student and those are likely to be highly correlated because high ability students are likely produce high scores in any domains, in general. Then, if we have such data as repeated

[1] Longford used the Fisher scoring algorithm, which can produce the estimate outside of the parameter space. The results are the case of this out-of-bounds estimate by Fisher scoring.

measurements for students (testlets are nested within subjects), we can expect the high tau-correlation matrix for the random coefficients that describes between-student variation.

If we take down our analysis down to the item level in the above scenario, then we have dichotomous outcome instead of continuous outcome. This is a scenario of using an item response theory (IRT) model (Hambleton, Swaminathan, & Rogers (1991)). The IRT model fits a nonlinear function that decomposes the log-odds of a correct response into an item-specific part and a person-specific part. The most standard IRT model is a unidimensional model and is in an exploratory factor analysis mode, but there is a multi-dimensional IRT model in terms of dimensionality, for example, see Reckase (1991). Even the multi-dimensional IRT model is still estimated by exploratory mode analysis, i.e., we are interested in how many dimensions we need to represent the item-person interaction. It is possible to execute a confirmatory mode analysis if we formulate the multi-dimensional IRT model with a hierarchical-model format. We will illustrate this point in Chapter 6. In this sense, the IRT model can be considered to be a non-linear version of correlated outcomes within subjects, corresponding to the previous linear and continuous version of correlated outcomes within subjects. In fact, Rasch model, the simplest IRT model, can be formulated by a hierarchical model by conceptualizing that items nested within examinee (Kamata (1998)). Raudenbush & Sampson (1999) applied the Rasch model to the items measuring neighborhood environments in terms of physical and social disorder. Their model can be seen as a multivariate Rasch model because they used dummy variables to represent different constructs.

Most of the above cases involve micro aspects where we are interested in individual differences and in individual variability. However, we can find the examples in other substantive disciplines. For example, from organizational sociology, data from National Adult Literacy Survey study analyzed by Raudenbush and Kasim (1998) involves the regression of literacy on ethnicity conceived as European-American, African-American, Hispanic-American, Asian-American, and Other ethnicity. Thus, four dummy variables are required to represent ethnicity. If these dummy variables' regression coefficients are allowed to vary from state to state, it can be speculated that those random coefficients might be highly correlated because of the similar social distribution of the minority disadvantage. Another example can be found in meta-analysis literature. Becker (1988) found the rather high correlation 0.91 between the experimental and control standardized mean changes across studies.

These above cases all suggest that those random coefficients share some part in common and tell us that once we know one of the random errors, we can discern the other to some extent. Then, it might be reasonable to think about a latent variable that is shared by those random coefficients and that produced the high correlation. In this scenario, a factor analysis model is one of the candidates to represent this idea. In factor analysis, we consider a small number of latent variables that explain the correlation among a larger number of observed variables. In this sense, the factor analysis model is useful for obtaining a parsimonious summary by reducing the number of parameters in a technical sense, but also it might extract the essential relationship among the constructs that play a

central role in the social theory. Further, it might help explore and elaborate or test and confirm the theory in mind.

Also, there are computational advantages of incorporating the factor analysis model to HLM. That is, in the current HLM, when some of the correlations get close to the boundary values such as 1 or -1, then its convergence gets very slow, or in the worst scenario, it terminates without giving the estimates. If we use the factor analysis model, we can expect quick and stable convergence because in factor analysis model, we are not estimating the covariances that are close to their boundary.

1-2. Overview of Past Studies of Multilevel Latent Variable Modeling

In this section, I review the current state of the research in the field of multilevel latent variable modeling, which is an effort in combining multilevel modeling, structural equation modeling, and item response theory (IRT) modeling. It is a goal of scientific research to integrate different classes of models.

Since the original model was developed in each tradition to solve specific types of problems, each modeling framework inherently has its own characteristic strengths and weaknesses. If we take the approach of incorporating another framework based on one framework, then the newly-developed model and the methodology are reminiscent of the parent's model characteristics.

Therefore, here we review those studies by emphasizing the underlying ideas to develop the model and summarize the similarities and differences as well as the strengths

and weaknesses. We also describe a methodology that can be a potential building block of a broader class of model, specifically, that of Jennrich & Schluchter's (1986) work.

1-2-1. Papers on Latent variables in HLM

Incorporating latent variables within the hierarchical linear model (HLM) can be found in Raudenbush, Rowan, and Kang (1991), where classical test theory was introduced into their level-1 model to handle measurement error variation. Most social science research involves instruments that are supposed to measure certain constructs, but they produce measurement errors. If these measurement errors are ignored, then maximum likelihood estimates of correlations will be attenuated (Lord & Novick, 1968, page 69-74) as will be the regression coefficients. As a result, we will get biased estimates. The instrument used by Raudenbush, Rowan, and Kang (1991) is the Administrator Teacher Survey (ATS), which is a questionnaire that asked the high school teachers about school climate consisting of 35 items that are supposed to measure the five constructs. They used a three-level model in which the level-1 units were items, the level-2 units were teachers, and the level-3 units were schools. In their model, using five dummy variables at level-1 specified which items were intended to measure which of five constructs. Thus five latent variables which played a role of true scores in classical test theory were specified in the level-1 model. At level-2, those latent variables in turn became multivariate outcome variables defined on teachers. At level-3, the school mean of each entry varied over schools. The unrestricted model had no predictors in each level. This enabled the authors to do psychometric analysis, i.e., to estimate the internal

consistency reliability of observed-scale scores. A major contribution of this approach was that they clarified the ambiguity that occurred from computing usual Cronbach's α (1951) by ignoring the school-clustering effects and separated it to two components, the reliability of the teacher-level measures and the reliability of school-level measures. The results showed that we could recover the attenuated correlation by incorporating the measurement model. Another aspect of this level-1 model can be considered as a special case of confirmatory factor analysis with factor loading weights specified as unity. Actually they found evidence that the number of factors was less than five by computing the eigenvalues for each teacher or school-level covariance matrices. This paper incorporated the measurement error in dependent variables in their level-1 model which served as a measurement model, and then used the true scores as dependent variables at level-2 that served as a structural model. In this sense, this was a prelude for complete structural equation modeling (SEM) which incorporate measurement errors in both dependent and independent variables.

In fact, Raudenbush and Sampson (1997) further extended this line of approach to formulate a model that allowed measurement errors both in dependent and independent variables. The model was formulated to examine the extent to which neighborhood social control (Z) mediates the association between neighborhood social composition (X) and violence (Y) in Chicago. The model was three-level model (measure, person, and neighborhood were the units for each level), and Y and Z involved measurement errors and X did not. Also Y and Z were measured at level-1 and X was at level-3. Then at level-1, for measure i for person j in neighborhood k, a measurement model was

formulated to represent the set of true scores and error scores using dummy variables as in the previous Raudenbush, Rowan, and Kang (1991)'s model. That is:

Level-1:

$$R_{ijk} = D_{1ijk}(Y_{jk} + \varepsilon_{jk}) + D_{2ijk}(Z_{jk} + \nu_{jk}), \qquad (1\text{-}1\text{-}1)$$

$$\varepsilon_{jk} \sim N(0, \sigma^2_{1jk})$$

$$\nu_{jk} \sim N(0, \sigma^2_{2jk}),$$

where D_{1ijk} is an indicator variable taking on a value of 1 if R_{ijk} is a measure of perceived violence and 0 if not, and D_{2ijk} is an indicator variable taking on a value of 1 if R_{ijk} is a measure of social control and 0 if not (i = 1,2; j = 1,...,J_k; k = 1,...,K). Note that this formulation allows missing data such as when a person provides a measure of perceived violence but not of social control, and vice versa. The level-2 model was formulated for person j in neighborhood k:

Level-2:

$$Y_{jk} = Y_k + W^T_{jk}\pi_{yk} + r_{yjk}$$

$$Z_{jk} = Z_k + W^T_{jk}\pi_{zk} + r_{zjk} \qquad (1\text{-}1\text{-}2)$$

$$\begin{pmatrix} r_{yjk} \\ r_{zjk} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Omega_{yy} & \Omega_{yz} \\ \Omega_{zy} & \Omega_{zz} \end{pmatrix}\right),$$

where $W^T_{jk} = (Age_{jk}, Gender_{jk}, SES_{jk})$ that was used for adjusting for background differences. The level-3 model describes the variation across neighborhood of adjusted neighborhood mean perceived violence and social control:

Level-3:

$$Y_k = X_k^T \beta_y + u_{yk},$$

$$Z_k = X_k^T \beta_z + u_{zk}, \qquad (1\text{-}1\text{-}3)$$

$$\begin{pmatrix} u_{yjk} \\ u_{zjk} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} T_{yy} & T_{yz} \\ T_{zy} & T_{zz} \end{pmatrix}),$$

where X_k^T was measures about neighborhood environments such as poverty concentration, ethnic isolation, and percent of foreign born, that were considered measured without error. For other regression components in level-2 model was fixed for parsimony:

$$\pi_{yk} = \beta_{y0}, \qquad (1\text{-}1\text{-}4)$$

$$\pi_{zk} = \beta_{z0}.$$

Since the above model has the same structure of the model in Raudenbush, Rowan, and Kang (1991), the parameters such as β_y, β_z, β_{y0}, β_{z0} for fixed effects, and T_{yy}, T_{yz}, and T_{zz} for the random effects could be estimated by the standard method. However, their interest was to evaluate the mediating effect of social control (Z) on the regression model of social composition (X) on perceived violence (Y) at the neighborhood level. This was done by considering conditional distribution of Y|Z, X from Y,Z|X . Since $(Y_k^T, Z_k^T)^T$ in the model Equation 1-1-3 had multivariate normal distribution, the conditional distribution of Y|Z, X was obtained by standard normal theory. To obtain the decomposition formula, the standard structural equation modeling (SEM) (or originally, the path analysis) idea was used. That is, they first wrote the

regression of Y on X and Z, and the regression of Z on X. Creating the reduced form (the regression of Y on X) gave the formulae for the decomposition of total effect into direct and indirect effects. The all necessary quantities to evaluate the decomposition were the functions of β_y, β_z, β_{y0}, β_{z0} T_{yy}, T_{yz}, and T_{zz}.

The decomposition of total effects into direct and indirect effect among latent variables is often executed by the SEM and thus it can be considered to be an attempt to integrate the HLM and the SEM to a broader class of model basing on HLM and then incorporating SEM. The only difference is that the model by the HLM used classical test theory model instead of factor analysis model that is often formulated by the SEM.

If the dependent variable is not continuous, say, dichotomous, we need a different model and the estimation method. Often, measurement instruments in social sciences such as cognitive achievement test consists of dichotomous items. In Raudenbush and Sampson (1999), the systematic social observation (hereafter, SSO) was used to measure the physical and social disorders of face blocks in Chicago neighborhood. The evaluation of the SSO's items were dichotomous. A three-level model involved items within face blocks within neighborhood. The level-1 model was a Rasch model (Rasch, 1960) that related item responses to item difficulties and face-block severities. At level-2, between face-blocks within cluster, the face blocks were outcomes depending on neighborhood - level intercepts. At level-3, neighborhood intercepts varied over neighborhoods. This analysis produced psychometric properties such as internal consistency reliabilities at each level.

Similarly, Cheong and Raudenbush (1999) applied Rasch Model to Child behavior Check List 4-18 (hereafter CBCL 4/18) (Achenbach, 1991) to calibrate the extent of severity of the externalizing behavioral problems of the children for each item. They used dummy variables to represent two different constructs (aggression and delinquency) indicator in level-1.

In psychometrics, the application of HLM to Item Response Theory (IRT) model can be found in Kamata (1998), where he used the Rasch Model. Note that in psychometric literature, the item severity to measure the extent to which the neighborhood was disordered is replaced by item difficulty and the neighborhood propensity to social disorder is referred to as person ability. Bock (1988), and Adams, Wilson & Wu (1997) took a different direction, where they started from a 2 or 3-parameter IRT model.

1-2-2. Structural Equation Models (Mean and Covariance Structure Models)

As we mentioned in the above when we refer to Raudenbush and Sampson (1997), the framework of separating a statistical model into a measurement model and structural model has a long tradition in mean and covariance structure models or structural equation models (SEM), which are implemented by package softwares such as LISREL (Jöreskog & Sörbon (1995), EQS (Bentler & Wu (1995)), AMOS (Arbuckle (1995)), and SAS CALIS procedure (SAS Institute Inc. (1990)), and M-Plus (Muthén (1998)) among others. In mean and covariance structure models, the measurement model is usually expressed as a confirmatory factor analysis model (CFA), which utilizes factors

as latent variables. The primary goal of factor analysis is to explain the covariances or correlations between many observed variables by relatively few underlying unobserved factors. In this sense, it is a data reduction technique. CFA, as contrasted to traditional exploratory factor analysis (EFA), implies that a model is constructed in advance, the number of factors is set by the researcher, whether a factor influences an observed variable is specified, some direct effect of factor on observed variables are fixed to zero or some other constant, covariances of factors can be estimated or set to any value, etc. Thus, in CFA, there are more opportunities that the researcher's idea are reflected on the model.

As suggested in the previous section, hierarchical linear model (HLM) and structural equation modeling (SEM) has much in common as methodology. In growth modeling context, Willet & Sayers (1994) showed that growth trajectory approach taken by HLM could be done by SEM if the data are balanced. Several methodologists claim that, compared to HLM, SEM approach has an advantage in terms of flexibility of modeling of the complex structure of variance-covariance matrix that naturally arises when we formulate a complicated causal model (Muthén & Curran (1997), Willet & Sayer (1996), Chou & Bentler (1998)). However, it is limited in dealing with nested structure of data and unbalanced designs because the model is estimated from sufficient statistics such as either the correlations or the means and the covariances. An attempt to incorporate clustered data structure and dealing with unbalanced and missing data which is a feature advantage of HLM can be found in, for example, Muthén (1989).

Recently Muthén (1998) developed a methodology which integrates categorical and continuous latent variable models. Latent categorical variable plays a role of a classification variable in multiple-group SEM, but the group membership is unobserved and is determined from the data such that a person is classified into the class that has the highest probability. In growth modeling context, this methodology not only adds more flexibility of the modeling that can reflect a class of substantive developmental theories which concerns with discrete transitions or qualitative changes rather than quantitative changes, but also provides a practically very useful way of predictive diagnostics in preventive or intervention research.

Comprehensive and systematic treatment of latent variable models including factor analysis, latent trait analysis (IRT), latent profile analysis, and latent class analysis can be found in Bartholomew & Knott (1999).

1-2-3. Modeling the Error Structure

An attempt to incorporate various patterns of the error covariance structure into unbalanced repeated measures was made by Jennrich and Schluchter (1986) in a longitudinal study context, where we can consider the case when the study was executed perfectly in a designed way, i.e., no missing observations. If there are no missing observation, we can utilize the standard MANOVA, or MANCOVA model. The key idea of their model was to introduce a covariance structure model for the complete data and then to link the observed outcome and the complete data by missing observation indicators. They formulated the standard MANCOVA growth model for that complete

data and considered various error structures for the complete data error terms that were used in the standard models such as time series model. Those include compound symmetry, heteroskedastic errors, auto-regressive errors, and so on. Exploratory mode factor analysis model which put a constraint on covariance matrix Φ, $\Phi = \mathbf{I}$ (identity matrix) was one of the structures they listed. The key assumption for the missing data matrix was missing at random (MAR).

This line of approach can be seen in Thum (1997). He explicitly stated that the key assumption for the missing data matrix was missing at random (MAR). Thum specially focused on individual differences and individual variation. Starting from standard MANCOVA model for repeated measurement on human subjects, he formulated the two-level hierarchical linear model. In addition to showing various patterns of covariance structure of both level-1 and level-2, he addressed the sensitivity of inferences for small sample size data that is often the case in psychological studies by using a multivariate t-distribution instead of a multivariate normal distribution for modeling the variation of random coefficients.

1-2-4. Multilevel Factor Analysis

Longford and Muthén (1992) analyzed data having eight domains in mathematics for the 8th grade from the Second International Mathematics Study (SIMS) carried out in 1982, and they considered a model with factor structures at both the student and school level. They used Fisher scoring to obtain maximum likelihood estimates for this model. Their theory was developed for the exploratory factor analysis phase, not for the

confirmatory phase. The main objective was to see whether the factor loading matrices had the same patterns at the student and school level.

Muthén (1991) developed the multilevel factor analysis model from the SEM perspective and showed that if sample size per the level-2 unit was large, the conventional SEM approximate estimator worked well, providing similar results to those of the Longford and Muthén (1992) for the same SIMS data set.

Muthén, Khoo, and Gustafsson (1998) extended this approach to multiple groups. They used the eighth graders' 16 achievement scores from National Educational Longitudinal Study (NELS) administered in 1988, conceptualizing that urban Catholic and urban public schools are two distinct populations. Thus this methodology is a generalization of conventional latent variable multiple-group analysis to two-level clustered data.

Though multilevel factor analysis is already developed, there are certain cases where it appears useful to incorporate the factor analysis model directly into standard hierarchical linear and non-linear models. That is, seeing the level-1 model of HLM as a model that generates latent variables that represent the true scores , in a broader sense, we apply the factor model for those random parameters such that the true score can be decomposed into a common score and a specific score in level-2 model. Then, we can clearly assess how much variation is explained by communality and how much is by the specificity. This decomposition can not be done by standard factor analysis because specificity is confounded with error variance in the model. Thus, in Chapter 2, we will give a rationale for incorporating a factor structure in level-2 in HLM in more detail.

Chapter 2. Specific Settings in which the Proposed Models are Useful

In this chapter, contexts in which incorporating factor analysis structures into HLM would be useful will be discussed in more detail. Next the model can be laid out in terms of a hierarchical linear model. Technical issues such as method of estimation and derivation of computational formulae will be discussed in Chapter 3.

2-1. Slopes-as-Outcomes Model in Organizational Studies

Suppose that in a school effectiveness study such as High School and Beyond (Coleman, Hoffer, & Kilgore (1982)), we want to know that how much the relationship between math achievement and student characteristics such as race, gender, SES vary among schools. Suppose, in fact, that there are six covariates at level 1, then naturally, we might formulate

L-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + \beta_{4j}x_{4ij} + \beta_{5j}x_{5ij} + \beta_{6j}x_{6ij} + \varepsilon_{ij}, \qquad (2\text{-}1\text{-}1)$$

where $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$ and $x_{1ij}, x_{2ij}, x_{3ij}, x_{4ij}, x_{5ij}$, and x_{6ij} is the certain covariates of student characteristics for student i in school j.

At level 2, suppose that those coefficinets all randomly vary among schools after being accounted for by a certain school characteristic W_j.

L-2:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01} W_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} W_j + u_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21} W_j + u_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31} W_j + u_{3j} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41} W_j + u_{4j} \\
\beta_{5j} &= \gamma_{50} + \gamma_{51} W_j + u_{5j} \\
\beta_{6j} &= \gamma_{60} + \gamma_{61} W_j + u_{6j}
\end{aligned} \tag{2-1-2}$$

where $u_j = (u_{0j}, u_{1j}, u_{2j}, u_{3j}, u_{5j}, u_{6j})^T$ is considered to be distributed as mean of $\mathbf{0}$ and the covariance of τ, a 7×7 symmetric matrix. This is fairly large model since the number of unique parameters in tau-matrix is $(7 \times 8)/2=28$ and the data may not provide enough information to detect all the variances and covariances, or the current algorithm may not converge in reasonable time. However, if indeed u_{1j} and u_{2j} go together and u_{3j}, u_{4j}, u_{5j} and u_{6j} go together, then we formulate the factor model for $u_j = (u_{0j}, u_{1j}, u_{2j}, u_{3j}, u_{4j}, u_{5j}, u_{6j})^T$ as

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \\ u_{6j} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_2 \\ 0 & 0 & \lambda_3 \\ 0 & 0 & \lambda_4 \end{pmatrix} \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \\ \eta_{2j} \end{pmatrix}, \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \\ \eta_{2j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{01} & \psi_{02} \\ \psi_{10} & \psi_{11} & \psi_{12} \\ \psi_{20} & \psi_{21} & \psi_{22} \end{pmatrix} \right). \tag{2-1-3}$$

In matrix notation, the above models can be written as

L-1:

$$Y_j = X_j \beta_j + \varepsilon_j, \ \varepsilon_i \sim N(0, \sigma^2 \mathbf{I}_{n_j}) \tag{2-1-4}$$

L-2:

$$\beta_j = W_j\gamma + \Lambda\eta_j. \tag{2-1-5}$$

Note that

$$\tau = D(u_j) = \Lambda\Psi\Lambda^T. \tag{2-1-6}$$

By using a factor model, we reduce the number of variance-covariance parameters estimated from 28 to 10.

2-2. Growth Models in Human Development Studies

Suppose that in a child development study, an outcome is modeled by a linear function of age and want to know how much variability exists among children for intercept and the rate of growth. Then, the model is

L-1:

$$Y_{ti} = \pi_{0i} + \pi_{1i}a_{ti} + \varepsilon_{ti},\ \varepsilon_{ti} \sim N(0,\sigma^2) \tag{2-2-1}$$

where a_{ti} is the age of child i at time t for $t = 1,..., T_i$, and $i = 1,..., n$. Or, in more compactly,

$$Y_{ti} = A_{ti}^T\pi_i + \varepsilon_{ti}, \tag{2-2-2}$$

where $A_{ti} = (1, a_{ti})^T$ and $\pi_i = \left(\pi_{0i},\ \ \pi_{1i}\right)^T$.

In matrix notation for child i,

$$Y_i = A_i\pi_i + \varepsilon_i,\ \varepsilon_i \sim N(0,\sigma^2\mathbf{I}_{T_i}) \tag{2-2-3}$$

where $Y_i = (Y_{1i},...,Y_{T_i i})^T$, $A_i = \begin{pmatrix} A_{1i}^T \\ \vdots \\ A_{T_i i}^T \end{pmatrix}$, and $\varepsilon_i = (\varepsilon_{1i},...,\varepsilon_{T_i i})^T$. $\mathbf{I}_{T_i}$ is the $T_i \times T_i$ identity matrix.

The level-2 model describes the variability of those person-specific parameters among children.

L-2:

$$\begin{matrix} \pi_{0i} = \gamma_{00} + u_{0i} \\ \pi_{1i} = \gamma_{10} + u_{1i} \end{matrix}, \begin{pmatrix} u_{0i} \\ u_{1i} \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}). \tag{2-2-4}$$

Or, in matrix notation,

$$\pi_i = \gamma + u_i, u_i \sim N(0, \tau). \tag{2-2-5}$$

Now suppose we suspect that u_{0i} and u_{1i} are correlated with the correlation of 1 or -1. Then using the proposed model, we formulate

$$\begin{matrix} u_{0j} = \eta_{0j} \\ u_{1j} = \lambda \eta_{0j} \end{matrix}, \ \eta_{0j} \sim N(0, \psi_{00}). \tag{2-2-6}$$

In matrix notation,

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} 1 \\ \lambda \end{pmatrix} (\eta_{0j}), \ \eta_{0j} \sim N(0, \psi_{00}).$$

or

$$u_j = \Lambda \eta_j, \ \eta_{0j} \sim N(0, \Psi). \tag{2-2-7}$$

This implies that the variance and covariance matrix of u_j is

$$\tau = \Lambda \Psi \Lambda^T. \tag{2-2-8}$$

Thus, using a factor model, we reduce the number of parameters τ from 3 ($\tau_{00}, \tau_{10}, \tau_{11}$) to 2 (ψ_{00}, λ).

2-3. Psychometric Analysis of Multivariate Outcomes

Suppose that we have a math test for 3rd grade comprised of 9 testlets (a testlet is a group of questions that are closely related in terms of the topic within a test) and the score from each testlet is the number of correct responses aggregated from the items that belong to the testlet. In this setting, each score from a testlet can be considered to be continuous and the score is standardized so that each testlet score has approximately the same variance, or the same standard deviation in the population, say, a unit standard deviation. This homogeneity of variance assumption is not restrictive for application, because test scores are inherently interval scale so that we can always standardize as long as we only consider one population problem that doesn't involve group comparisons. Thus in practice we standardize the scores using sample standard deviations for each testlet among students. This standardization procedure should be acceptable if the number of students is fairly large so that sample standard deviation is very close to the population standard deviation. Further, suppose that the test was constructed so that the first three testlets measure arithmetic proficiency, the second four testlets measure algebraic proficiency, and the third two testlets measure geometric proficiency. Then, the following model might be formulated. For testlet score Y_{ij} for testlet i and student j,

L-1:

$$Y_{ij} = \beta_{1j} D_{1ij} + \beta_{2j} D_{2ij} + \beta_{3j} D_{3ij} + \varepsilon_{ij}, \qquad (2\text{-}3\text{-}1)$$

where D_{1ij}, D_{2ij} and D_{3ij} are the indicators of which construct the item i indicates, and $\varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$. More explicitly, if the examinee i has all the testlet scores, then

$$\begin{pmatrix} Y_{1j} \\ Y_{2j} \\ Y_{3j} \\ Y_{4j} \\ Y_{5j} \\ Y_{6j} \\ Y_{7j} \\ Y_{8j} \\ Y_{9j} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1j} \\ \varepsilon_{2j} \\ \varepsilon_{3j} \\ \varepsilon_{4j} \\ \varepsilon_{5j} \\ \varepsilon_{6j} \\ \varepsilon_{7j} \\ \varepsilon_{8j} \\ \varepsilon_{9j} \end{pmatrix}. \qquad (2\text{-}3\text{-}2)$$

Note that the level-1 model is actually a classical test theory model that decomposes the observed score Y_{ij} into the true score and the error score ε_{ij}. In this example, there are three true scores for each student j, arithmetic proficiency true score β_{1j}, algebraic proficiency true score β_{2j}, and geometric proficiency true score β_{3j}, depending on which construct the item is supposed to measure. In matrix notation, Equation (2-3-2) can be written as

$$Y_j = X_j\beta_j + \varepsilon_j, \varepsilon_j \sim N(0,\sigma^2\mathbf{I}_{n_j}). \qquad (2\text{-}3\text{-}3)$$

Then, at level-2, the three true scores for the examinee j vary among examinees.

L-2:

$$\begin{matrix} \beta_{1j} = \gamma_{10} + u_{1j} \\ \beta_{2j} = \gamma_{20} + u_{2j} \\ \beta_{3j} = \gamma_{30} + u_{3j} \end{matrix}, \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{21} & \tau_{22} & \tau_{23} \\ \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix}\right). \qquad (2\text{-}3\text{-}4)$$

Or in matrix notation,

$$\beta_j = \gamma + u_j,\ u_j \sim N(0, \tau). \tag{2-3-5}$$

If we suspect the correlation between u_{1j} and u_{2j} to be near unity, then we formulate

$$\begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix},\ \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{00} & \psi_{01} \\ \psi_{10} & \psi_{11} \end{pmatrix} \right). \tag{2-3-6}$$

In matrix form, this is written as

$$u_j = \Lambda \eta_j,\ \eta_{0j} \sim N(0, \Psi) \tag{2-3-7}$$

and the variance-covariance matrix of u_j is

$$\tau = \Lambda \Psi \Lambda^T. \tag{2-3-8}$$

By using the factor model, we reduce the number of parameters in τ from 6 to 4. Also, notice that by using a factor model at level-2, we might be able to say that there are actually two constructs instead of three. But an alternateive argument is that, as the variances are different, i.e., $\tau_{22} \neq \tau_{33}$, we might consider that the second and the third measure the different construct (Cheong & Raudenbush (1999)). In either interpretation, the reliability (Cronbach's α type internal consistency) of the test score on the examinee level (level-2) can be obtained.

Chapter 3. Approach and Model Layouts

3-1. The Model

All the examples in sections 1, 2, and 3 of Chapter 2 can generally be written as L-1:

$$Y_j = X_j\beta_j + r_j, \tag{3-1}$$

where Y_j is a $n_j \times 1$ vector, X_j is a $n_j \times R$ matrix, β_j is a $R \times 1$ vector, and r_j is a $n_j \times 1$ vector, and

$$r_j \sim N(0, \sigma^2 \mathbf{I}_{n_j}) \tag{3-2}$$

where $\mathbf{I}_{n_j}$ denotes identity matrix of size n_j.

L-2:

$$\beta_j = W_j\gamma + u_j \tag{3-3}$$

where W_j is a $R \times F$ matrix for F is a number of fixed effects parameters, γ is a $F \times 1$ vector, u_j is a $R \times 1$ vector. And for u_j, we considered factor structure,

$$u_j = \Lambda\eta_j, \tag{3-4}$$

where Λ is a $R \times M$ matrix, η_j is a $M \times 1$ vector $(R \geq M)$, and

$$u_j \sim N(0, \tau), \tag{3-5}$$

$$\eta_j \sim N(0, \Psi). \tag{3-6}$$

Thus, we have

$$\tau = \Lambda\phi\Lambda^T. \tag{3-7}$$

The combined model of the level-1 and the level-2 models is

$$Y_j = X_j W_j \gamma + X_j \Lambda \eta_j + r_j . \tag{3-8}$$

Considering the case where we fix some of the elements of u_j and not others of β_j, the more general model can be written as

$$Y_j = A_{1j}\gamma + A_{2j}\Lambda\eta_j + r_j , \tag{3-9}$$

where

$$r_j \sim N(0, \sigma^2 \mathbf{I}_{\mathbf{n}_j}) \text{ and } \eta_j \sim N(0, \Psi). \tag{3-10}$$

Note that Y_j is a $n_j \times 1$ vector of observed scores, A_{1j} is a $n_j \times F$ matrix where F is a number of fixed effects parameters, γ is a $F \times 1$ parameter vector, A_{2j} is a $n_j \times R$ matrix for R is a number of random effects, Λ is a $R \times M$ ($R \geq M$) factor loading matrix, η_j is a $M \times 1$ vector of factor scores, and r_j is a $n_j \times 1$ residual vector. And σ^2 is a common variance of each element of r_j, $\mathbf{I}_{\mathbf{n}_j}$ is a $n_j \times n_j$ identity matrix, and Ψ is a variance-covariance matrix of η_j.

3-2. Similarities and the Difference from the Multilevel Factor Analysis Model

The multilevel factor analysis model formulated by Longford and Muthén (1992) was

L-1:

$$Y_{ij} = \mu_j + \Lambda_1 \eta_{1ij} + \xi_{ij} \tag{3-11}$$

L-2:

$$\mu_j = \mu + \Lambda_2 \eta_{2j} + \xi_{2j} . \tag{3-12}$$

Thus this model assumes each student i in school j has p-variate complete observation.

If we translate the above model to the HLM notation, we would first formulate an unconditional model (no predictors in both level-1 and level-2 model) as

L-1:

$$Y_{ij} = \beta_{0j} + r_{ij},\ r_{ij} \sim N_P(0, \Sigma_1), \tag{3-13}$$

where Y_{ij} is a $P \times 1$ vector of outcomes for student i in school j, β_{0j} a $P \times 1$ vector of the school means, r_{ij} a $P \times 1$ vector of student level error, and Σ_1 is a $P \times P$ covariance matrix that is common to all the students across schools, and that represents the within-school between-student variability for P variates. Now we consider the school level model.

L-2:

$$\beta_{0j} = \gamma_{00} + u_{0j},\ u_{0j} \sim N_P(0, \Sigma_2) \tag{3-14}$$

where γ_{00} is the vector grand means, and u_{0j} is the vector of school level disturbances, and Σ_2 is a $P \times P$ covariance matrix that represents the between-school variability. If we make the combined model by plugging Equation 3-14 into Equation 3-13, we obtain

Combined model:

$$Y_{ij} = \gamma_{00} + u_{0j} + r_{ij} \tag{3-15}$$

Let $\Sigma \equiv D[Y_{ij}]$, $P \times P$ matrix. Then from Equation 3-15, it is clear that

$$\Sigma = \Sigma_1 + \Sigma_2. \tag{3-16}$$

Now we incorporate factor analysis model both in level-1 and level-2 residuals. That is, for level-1, let

$$r_{ij} = \Lambda_1 \eta_{1ij} + \xi_{1ij},\ \eta_{1ij} \sim N(0, \Psi_1),\ \xi_{1ij} \sim N(0, \Omega_1) \tag{3-17}$$

where Λ_1 is a $P \times M_1$ level-1 factor loading matrix , η_{1ij} is a $M_1 \times 1$ level-1 factor score vector, ξ_{1ij} is a $P \times 1$ level-1 uniqueness vector, and Ω_1 is a diagonal matrix.

For level-2, let

$$u_j = \Lambda_2 \eta_{2j} + \xi_{2j},\ \eta_{2j} \sim N(0, \Psi_2),\ \xi_{2j} \sim N(0, \Omega_2) \tag{3-18}$$

where Λ_2 is a $P \times M_2$ level-2 factor loading matrix , η_{2j} is a $M_2 \times 1$ level-2 factor score vector, ξ_{2j} is a $P \times 1$ level-2 uniqueness vector, and Ω_2 is a diagonal matrix. After incorporating the factor structure, we have

$$\Sigma_1 = \Lambda_1 \Psi_1 \Lambda_1^T + \Omega_1 \tag{3-19}$$

and

$$\Sigma_2 = \Lambda_2 \Psi_2 \Lambda_2^T + \Omega_2 . \tag{3-20}$$

Note that since they were interested in exploring the factor pattern, they gave constraints $\Psi_1 = \mathbf{I}_{\mathbf{M}_1}$ and $\Psi_2 = \mathbf{I}_{\mathbf{M}_2}$ in order for the model to be identifiable. Thus, covariance matrices actually modeled are

$$\Sigma_1 = \Lambda_1 \Lambda_1^T + \Omega_1 \tag{3-21}$$

and

$$\Sigma_2 = \Lambda_2 \Lambda_2^T + \Omega_2 . \tag{3-22}$$

They were interested in testing

$$H_0: \Lambda_1 = \Lambda_2 .$$

This hypothesis involves testing $M_1 = M_2$, i.e., the number of factors in student level is equal to the number of factors in school level.

To summarize the similarities and differences between my model and Longford & Muthén's model (1992), it can be stated in the followings.

Similarities:

1. Both approaches use a factor analysis model in the multilevel model.

Differences:

1. My model only has factor structure at level-2, but Longford & Muthén's model has factor structure at both level-1 and level-2.

2. My model has independent variables in both at level-1 and level-2, but Longford & Muthén's model does not have them either at level-1 or level-2.

3. My model is a confirmatory factor analysis model, but Longford & Muthén's model is an exploratory factor analysis model. That is, my model has known and fixed elements in factor loading matrix Λ, but Longford & Muthén's model estimates all the elements in level-1 and level-2 factor loading matrices (Λ_1, Λ_2) and let Ψ_1 and Ψ_2, factor score covariance matrices in level-1 and level-2, be identity matrices in order for the model to be identified.

4. My model can handle level-1 observations missing at random while Longford & Muthén's model requires all the students to have complete observations in terms of outcome.

3-3. **Identification**

When we start using a factor analytic type of the model that involves a decomposition of the covariance matrix which is a case of the confirmatory factor analysis (CFA) model, we often encounter model identification problems. The interpretations of the parameter estimates are meaningless if we estimate the parameters of the unidentified model. Therefore, model identification must be established before we estimate the parameters in the model.

Model identification is concerned with whether the parameters of the model are uniquely estimable, assuming a sufficiently large sample. Issues of model identification in confirmatory factor analysis (CFA), which is a special case of SEM (see the model at the section 2-2 of Chapter 1), are detailed in, for example, Bollen (1989). Therefore, here we briefly describe what it is, what the problems are, and what is the current status of our knowledge on model identification of CFA.

In CFA, the population covariance matrix Σ is a function of a $R \times 1$ vector of parameters θ, which contains all of the unknown and unconstrained parameters of the model.

A CFA model can be written as

$$\mathbf{y} = \Lambda\eta + \xi, \tag{3-23}$$

where $\mathbf{y}$ is a $R \times 1$ vector of deviation scores from the sample mean of observed variables so that we have $E(\mathbf{y}) = \mathbf{0}$; η is a $M \times 1$ vector of common factors; Λ is a $R \times M$ matrix of factor loadings relating the observed $\mathbf{y}$'s to the latent η's; and ξ is a $R \times 1$ vector of residual or unique factor, which is a combined element of unique factor

and measurement error. Let Σ be the population covariance matrix of $\mathbf{y}$, i.e., $\Sigma = D[\mathbf{y}] = E(\mathbf{y}\mathbf{y}^{\mathrm{T}})$. Let Ψ be the covariance matrix of the η, i.e., $\Psi = D[\eta]$, and let Ω be the covariance matrix of ξ, i.e., $\Omega = D[\xi]$. We assume that the population means of η and ξ are zero and η and ξ are uncorrelated. That is, we assume $E(\eta) = \mathbf{0}$, $E(\xi) = \mathbf{0}$, and $Cov(\eta, \xi) = E[\eta\xi^T] = \mathbf{0}$. Then, we obtain the covariance equation,

$$\Sigma = \Lambda\Psi\Lambda^T + \Omega. \tag{3-24}$$

Thus Σ is represented as a function of the Λ, Ψ, and Ω. In the CFA model, some constraints are placed on Λ and/or Ψ based on the prior knowledge of the subject matter that researchers are studying. Let q_1 be the number of unknown parameters in Λ, q_2 be the number of the unknown parameters in Ψ, and q_3 be the number of unknown parameters in the Ω, and let q be the total number of unknown parameters in the right hand side of Equation (3-24). Then we have $q = q_1 + q_2 + q_3$. We represent all of these parameters as a $q \times 1$ vector θ.

A model is said to be identified if all parameters in the vector θ are identified. Thus a model is identified if $\Sigma = \Sigma(\theta_1) = \Sigma(\theta_2)$ implying that $\theta_1 = \theta_2$, where θ_1 and θ_2 are the parameter vectors that have any specific values. Note that θ_1 and θ_2 are different even if one of the corresponding elements are different. Negating the proposition, the model is not identified if there exist θ_1 and θ_2 such that $\theta_1 \neq \theta_2$ and $\Sigma = \Sigma(\theta_1) = \Sigma(\theta_2)$. Another way of saying this is that if the parameters in θ are solved uniquely by a function of the population parameters in Σ, then the model is identified because estimation involves using sample data to obtain estimates of population parameters. In

confirmatory factor analysis (CFA), this involves using the sample covariance matrix of the observed variables, called $\mathbf{S}$, which is sufficient for Σ under a balanced design and given our assumptions, to estimate the parameters in Λ, Ψ, and Ω.

The latter definition of identification suggests a method to demonstrate model identifiability. That is, if all parameters in θ are represented by unique functions of elements of Σ, then the model is identified. If not, the model is said to be unidentified. When there is only one way to express the parameters in θ, the model is said to be just identified. If there are several ways, then the model is said to be overidentified. In most cases, we try to formulate an overidentified model, because it provides an opportunity to test our hypothesis. Representing a model with a smaller number of parameters than the number of unique parameters in the population covariance matrix Σ reflects our attempt to explain or describe a complex phenomena by a relatively simple theory.

The argument of identifiability of a multilevel factor model can be made in the same way since τ can be seen as the counterpart of a population covariance matrix of observed variables in SEM, which is denoted as Σ. That is, if τ is the positive semi-definite, the identification condition follows those given by the confirmatory factor analysis (CFA) model. Thus, we first summarize the current status of our knowledge about identification of CFA model.

The trouble in assessing identifiability of CFA models is that not only we have not found necessary and sufficient conditions, but also we have not yet even found a relatively general, widely applicable sufficient condition. Currently there are a couple of necessary conditions known. They are useful, but the model that satisfies the necessary

conditions is not necessarily an identifiable model. It only gives a hope that can be identified. Therefore, a pitfall in which we could fall is a case in which a model satisfies the necessary conditions, but is not an identifiable model, and we run the software without knowing the fact and the software gave the estimate. In this case, the estimate is meaningless[1]. On the other hand, models that satisfy sufficient condition are guaranteed to be identified. If we know the sufficient condition, we are sure that the estimates obtained by running the software are meaningful. Therefore, a sufficient condition is stronger condition than a necessary condition.

One way to know whether the formulated model is identified is to algebraically solve the simultaneous nonlinear equations in terms of unknown parameters. If each parameter in θ can be represented uniquely by a function of the elements in Σ and by the pre-specified parameters, then the model is identified. This is the only necessary and sufficient condition of model identification of CFA models that we know so far. However, since the simultaneous nonlinear equations to be solved become very complex when the model gets large and since we may not know a systematic way of solving them, it is often a very difficult task to know whether the unique solution is obtainable or not. Therefore, though it is a necessary and sufficient condition, solving the simultaneous nonlinear equations is not considered to be a good way to demonstrate the model identification. Long (1979) characterized this problematic status as "proving that a model

---

[1] A remedy for this is that using the different starting values of the parameters, we see if we get the same estimates. This strengthens our confidence that the model is identified, but it is not still enough because there are infinitely many starting values in the parameter space and we can not test all of them.

is identified presents one of the greatest practical difficulties in using the confirmatory factor model" (page 36).

In terms of necessary conditions for identification of CFA, there are generally two necessary conditions available. Kline (1998) summarized those necessary conditions as:

1) the number of parameters to be estimated ≤ number of observed unique variances and covariances ($q \leq \frac{R(R+1)}{2}$). (3-25)

2) every factor must have a scale, i.e., unit variance of each factor should be defined.

The second condition can be achieved by either fixing each factor variance to unity or by setting an element of each column of the factor loading matrix Λ to a constant, usually 1, which sets one of the factor loading one for each factor.

Since identifiable models must satisfy these conditions and both conditions are easy to check, they are very useful for screening unidentified models. However, even after screening the unidentifiable models using these necessary conditions, many unidentifiable models remain, having different patterns of factor loadings.

Some works have been made concerning sufficient condition for identification of CFA models. For example, the two and three indicator rules and Bollen's slightly more general condition for the three indicator rule (Bollen, 1989) provide some methods. Davis (1993) extended these works and found the general sufficient conditions for identification of CFA with factor complexity of one, where the factor complexity is defined as the number of factors on which observed variables load at most, and factor complexity of one

means that each observed variable loads on one and only one latent variable. This condition allows that measurement errors be correlated. However, the scope of these sufficient conditions are limited because these conditions apply only to CFA model of factor complexity of one, a rather simple model. Thus we only can use these rules of sufficient conditions for simple factor loading models. These do not apply to factor models that have a cross-loading structure, which is defined as some of the observed variables load on more than one factor.

It seemed promising at first look for a sufficient condition for uniqueness of factor loading matrix Λ under rotation provided by Howe (1955) and Jöreskog (1979). Jöreskog (1979) provided a sufficient condition for the uniqueness (not for model identification) of Λ under rotation for the confirmatory factor analysis (CFA) model. The CFA model that Jöreskog (1979) used adds another assumption for ξ, that is, each element in ξ is uncorrelated with each other element. This additional assumption constrains the Ω to be a $R \times R$ diagonal matrix which contains only the variances of ξ in the main diagonal, with all off-diagonal elements fixed to 0. It is assumed that all diagonal elements of Ω are unknown parameters to be estimated.

The scales of the latent variables η can be determined either by the requirement that Ψ is a correlation matrix or that there is a fixed constant (usually set as 1) in each column of Λ. The remaining elements of Λ are either fixed to zero or are unknown parameters to be estimated.

The sufficient conditions for uniqueness (again, not for model identification) that was stated by Jöreskog (1979) are:

(a) Ψ is a symmetric positive definite matrix with $\text{diag}(\Psi) = \mathbf{I}_R$,

(b) Λ has at least n-1 fixed zeros in each column, and

(c) Λ_S has rank n-1, where Λ_S, $S = 1,2,\ldots,n$, is the submatrix of Λ consisting of the rows of Λ which have fixed zero elements in the S^{th} column. (3-26)

Several authors of books on confirmatory factor analysis (for example, see pages 43-44 of Long (1979)) confused a sufficient condition for uniqueness of Λ under rotation with a sufficient condition for identification. Bollen and Jöreskog (1985) clarified this misunderstanding by showing an example in which uniqueness of Λ under rotation does not guarantee the identification of the model. An example of an unidentified model they used was unidentified because the parameters in Ω needed to be estimated. The Λ and Ψ were unique under rotation if the parameters in Ω were known. However, since parameters in Ω parameters need to be estimated, then unidentification occurred. Therefore, if there are no parameters in Ω in the model, it seems that the condition that Howe (1950) and Jöreskog (1979) proposed is a sufficient condition for identification as long as the examples I'm going to use.

Now we consider the identification issue for HLM that involves a factor analysis structure. If we replace the τ matrix in HLM as Σ in CFA, the same thing can be said about the model identification. The requirement of τ as a covariance matrix of u_j is that it should be positive semi-definite matrix, the same as the condition required for Σ. The only difference is that the HLM factor model (see Equation (3-7)) does not involve the unique residual variance, Ω (see Equation (3-24)), though it can be easily added.

Therefore, we restate the conditions for identification for the HLM factor model, which are the same for identification of CFA models, with an additional necessary condition that is useful for screening unidentified models.

We first describe the simple necessary conditions to check. If these conditions are not met, the model will not be identified. Thus, we can use these conditions to screen the unidentified models at the first stage of model consideration.

Necessary conditions:

1. the number of parameters estimated $\leq$ the number of parameters in τ $(q \leq \frac{R(R+1)}{2})$.
2. every factor must have a scale, i.e., unit variance of each factor should be defined. (3-27)
3. M^2 independent constraints on Λ (or Ψ) must be given, where M is the number of factors.

The first two are the same necessary conditions in CFA model. Necessary condition 1 is simply states that the number of equations must be at least as many as the number of unknowns. Thus, we need at least the same number of independent equations as the number of unknown parameters to be estimated.

Necessary condition 2 states that the M scales are necessary for each factor in the $M \times 1$ vector, η because in order to calibrate the variance of a random variable in terms of variance of factor, the unit variance of the factor must be defined on the first hand.

The third necessary condition that I propose derives from the indeterminacy of the factor model. Given a positive semi-definite covariance matrix τ and a number of M factors which satisfy the equation $\tau = \Lambda\Psi\Lambda^T$. Let $\Lambda^* = \Lambda T$ and $\Psi^* = T^{-1}\Psi T^{T-1}$ for any m by m invertible matrix T. Then $\Lambda^*\Psi^*\Lambda^{*T} = (\Lambda T)(T^{-1}\Psi T^{T-1})(T^T\Lambda^T) = \Lambda\Psi\Lambda^T = \tau$. Thus the pair (Λ^*, Ψ^*) produces the same τ as that of (Λ, Ψ). Therefore, in order to uniquely identify Λ and Ψ, we need to specify T, i.e., we need to impose M^2 (the number of elements in the matrix T) constraints on Λ or Ψ. Thus, if there are less than M^2 restrictions on either Λ or Ψ, then both Λ and Ψ are not be identifiable (See Pages 553 - 557 of Anderson (1984)).

There are two ways of determining the scales of each factor. One is to set the diagonal of Ψ to unity, which scales the variances of all latent factors to one. Another way is to set one of the elements in Λ for each column to one. Since we want to keep the same metric in η_j as in u_j in our multilevel confirmatory type factor model in Equation (3-9), we choose the second option. That is, we fix an element of each column of the factor loading matrix Λ as 1. By doing that, we are specifying a unique variance at level-2 unit as the reference variance of the factor.

If we choose to scale the latent factors η by setting an element for each column of Λ to one, then the 2nd and the 3rd necessary conditions can be summarized as one necessary condition, which is, the linear transformation matrix should be the identity matrix, i.e., $T = \mathbf{I_M}$. In this condition, the 2nd necessary condition is represented by the

fact that we fix all the diagonal elements of T to 1, and the 3rd necessary condition is represented by the fact that we set all the M^2 elements in T to either 1 or 0.

Though only two conditions out of three are independent, I formalize three necessary conditions for identification for our HLM2F model represented in Equation (3-9) for practical purpose in the followings:

If the model in Equation (3-9) can be identified, it must satisfy the following two conditions, otherwise the model cannot be identified:

1. the number of parameters estimated must be smaller or equal to the number of unique parameters in τ, i.e., $q \leq \frac{R(R+1)}{2}$.
2. every factor must have a scale, i.e., unit variance of each factor should be defined. (3-28)
3. the linear transformation matrix T must be the identity matrix, i.e., $T = \mathbf{I_M}$ under the transformations $\Lambda^* = \Lambda T$ and $\Psi^* = T^{-1}\Psi T^{T-1}$.

We confirm that the 3rd condition above is truly the necessary condition by the following reasoning: The model is not identified if the model is not unique under linear transformation of Λ and Ψ because there exist at least two pairs, (Λ, Ψ) and (Λ^*, Ψ^*) that produces the same τ, which is the definition of unidentified model. Therefore, our statement is if there exists a non-trivial $M \times M$ linear transformation matrix T, which means that T is not the identity matrix, then the model is not identified. Taking the contrapositive, we can say that if the model is identified, then the linear transformation matrix T must be the identity matrix. This states that $T = \mathbf{I_M}$ is a necessary condition for

the model identification. Note that the second necessary condition is exact when we choose to scale the latent factors η by setting an element in each column of Λ to one. If we choose another scaling option, which is setting $\Psi = \mathbf{I}_M$, then T can be either $T = \mathbf{I}_M$ or $T = -\mathbf{I}_M$ because of the arbitrariness of the sign of the factor loading matrix.

In terms of the sufficient conditions, we do not know them yet. We know that if we can solve the covariance equations uniquely for the unknown parameters in θ, the model is identified, and thus the solvability of covariance equations in terms of θ is a necessary and sufficient condition. In this situation, it seems most practical to proceed as follows:

1) Formulate the specific model and check its identification by two necessary conditions.
2) If the necessary conditions are satisfied, then we check the identification by attempting to algebraically solve the covariance equations. If it is solved, then the model is identified. If not, treat the model as unidentified.

Therefore, I present the model identification issue in detail for the specific models that will be used in Chapter 5. First, I will describe the context of the data because specification of the Λ matrix not only depends on the identifiability of the model but also on the substantive knowledge of the context. That is, it is not useful to estimate the parameters of the model that aren't substantively interesting.

As a classical application of a factor model, we consider a setting where subtests are nested within examinees. Though this does not require a hierarchical model, we can

see a hierarchical model as a technique for data reduction. This is useful because it allows comparison to the models where we have a lot of experience.

Specifically, suppose that a test involves 4 subtests, math1, math2, verbal1, and verbal2 and that each subject takes two parallel alternative forms at each test in a relatively short time span. We assume that there are no missing cases and that there is no learning effect or memory effect for the two occasions of the measurements. There are 100 subjects. Thus we create a situation in which subtests are nested within students and each student has 8 observations, which results in 800 observations total. Now we assume that the true scores of subtests math1 and math2 are perfectly correlated as are those for verbal1 and verbal2. Thus we can have 2 factors, mathematical and verbal proficiency. The correlation between mathematical and verbal true scores is 0.50. Then the model can be formulated in a hierarchical model form as

L-1:

$$\begin{aligned} Y_{ij} &= \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \varepsilon_{ij} \\ &= \sum_{q=1}^{4} \beta_{qj} X_{qij} + \varepsilon_{ij} \qquad , \ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2) \end{aligned} \tag{3-29}$$

where X_{qij} = 1 if ith observation is qth subtest, and 0 if not. The standard HLM2 specifies the level 2 model as

L-2:

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \\ \beta_{3j} &= \gamma_{30} + u_{3j} \\ \beta_{4j} &= \gamma_{40} + u_{4j} \end{aligned} \tag{3-30}$$

where $\begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{12} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{13} & \tau_{23} & \tau_{33} & \tau_{34} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_{44} \end{pmatrix}\right)$, and the supposed level 2 model reduces

the number of parameters at level 2 as

L-2:

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + \eta_{1j} \\ \beta_{2j} &= \gamma_{20} + \lambda_1 \eta_{1j} \\ \beta_{3j} &= \gamma_{30} + \eta_{2j} \\ \beta_{4j} &= \gamma_{40} + \lambda_2 \eta_{2j} \end{aligned} \tag{3-31}$$

where the unique four level 2 errors are reduced to two, i.e.,

$$\begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix}, \text{ and } \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix}\right). \tag{3-32}$$

Note that there is a relationship between u_j and η_j as

$$u_j = \Lambda \eta_j \tag{3-33}$$

and thus

$$\tau = \Lambda \Psi \Lambda^T. \tag{3-34}$$

In element wise, it can be represented in a set of equalities:

$$\tau = \Lambda \Psi \Lambda^T = \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{12} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{13} & \tau_{23} & \tau_{33} & \tau_{34} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_{44} \end{pmatrix} = \begin{pmatrix} \psi_{11} & \lambda_1\psi_{11} & \psi_{12} & \lambda_2\psi_{12} \\ \lambda_1\psi_{11} & \lambda_1^2\psi_{11} & \lambda_1\psi_{12} & \lambda_1\lambda_2\psi_{12} \\ \psi_{12} & \lambda_1\psi_{12} & \psi_{22} & \lambda_2\psi_{22} \\ \lambda_2\psi_{12} & \lambda_1\lambda_2\psi_{12} & \lambda_2\psi_{22} & \lambda_2^2\psi_{22} \end{pmatrix}, \tag{3-35}$$

and if we write this in a correlation form denoted as ρ_τ, then

$$\rho_\tau = \begin{pmatrix} 1.0 & 1.0 & \rho & \rho \\ 1.0 & 1.0 & \rho & \rho \\ \rho & \rho & 1.0 & 1.0 \\ \rho & \rho & 1.0 & 1.0 \end{pmatrix}, \tag{3-36}$$

where $\rho = corr(\eta_{1j}, \eta_{2j}) = \dfrac{\psi_{12}}{\sqrt{\psi_{11}}\sqrt{\psi_{22}}}$.

We set $\gamma_{10} = \gamma_{20} = \gamma_{30} = \gamma_{40} = 500$, $\sigma^2 = 25$, $\lambda_1 = 0.8$ $\lambda_2 = 1.2$, $\psi_{11} = 100$, $\psi_{22} = 100$, $\psi_{12} = 50$ so that $corr(\eta_{1j}, \eta_{2j}) = 0.50$, a medium size correlation. Thus we specified the τ as

$$\tau = \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{12} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{13} & \tau_{23} & \tau_{33} & \tau_{34} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_{44} \end{pmatrix} = \begin{pmatrix} 100 & 80 & 50 & 60 \\ 80 & 64 & 40 & 96 \\ 50 & 40 & 100 & 120 \\ 60 & 96 & 120 & 144 \end{pmatrix}, \tag{3-37}$$

and as a correlation,

$$\rho_\tau = \begin{pmatrix} 1.0 & 1.0 & 0.50 & 0.50 \\ 1.0 & 1.0 & 0.50 & 0.50 \\ 0.50 & 0.50 & 1.0 & 1.0 \\ 0.50 & 0.50 & 1.0 & 1.0 \end{pmatrix}. \tag{3-38}$$

The true model, as shown in the above, and from which we will generate the data is called "model 1".

Model 1: The assumed correct model

It has the factor loading matrix of

$$\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix}.$$

We see from it that the unique element u_{1j} and u_{2j} only loads on the first factor and the variance ratio is one to λ_1^2. Similarly, the unique element u_{3j} and u_{4j} loads only on the second factor and the variance ratio is one to λ_2^2.

The most general model among two factors can be formulated by letting all the elements in the factor loading matrix Λ be free, unknown parameters except setting two 1's at the position of (1,1) and (3,2) for scaling purposes. It is named "model 2" and is represented as:

Model 2: Saturated, but not identified model

$$\Lambda = \begin{pmatrix} 1 & \lambda_6 \\ \lambda_1 & \lambda_4 \\ \lambda_5 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix}$$

This is an unidentified model because it does not satisfy the necessary condition (2). This can be shown by checking the uniqueness of the linear transformation matrix T. Let

$$T = \begin{pmatrix} x & y \\ u & v \end{pmatrix}. \text{ Then, } \Lambda^* = \Lambda T = \begin{pmatrix} 1 & \lambda_6 \\ \lambda_1 & \lambda_4 \\ \lambda_5 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix} \begin{pmatrix} x & y \\ u & v \end{pmatrix} = \begin{pmatrix} x + \lambda_6 u & y + \lambda_6 v \\ \lambda_1 x + \lambda_4 u & \lambda_1 y + \lambda_4 v \\ \lambda_5 x + u & \lambda_5 y + v \\ \lambda_3 x + \lambda_2 u & \lambda_3 y + \lambda_2 v \end{pmatrix} = \begin{pmatrix} 1 & \lambda_6^* \\ \lambda_1^* & \lambda_4^* \\ \lambda_5^* & 1 \\ \lambda_3^* & \lambda_2^* \end{pmatrix}.$$

Thus we have conditions for Λ^* that need to be satisfied, which are

$$x + \lambda_6 u = 1,$$
$$\lambda_5 y + v = 1.$$

This simultaneous equations have infinitely many solutions other than $x = 1, y = 0, u = 1$, and $v = 1$, for example, $x = \frac{1}{2}, u = \frac{1}{2\lambda_6} v = \frac{1}{3}$ and $y = \frac{2}{3\lambda_5}$. Thus the linear transformation T is not the identity matrix. Therefore (Λ, Ψ) is not unique. Therefore model 2 is not identified.

Specifying the factor loading matrix as a 4 by 2 matrix reflects the researcher's belief about how many factors exist for the data. We specified that there are two factors and that $u_j = \left(u_{1j}, \quad u_{2j}, \quad u_{3j}, \quad u_{4j}\right)^T$ are linear combinations of two factors (η_{1j}, η_{2j}). This model is a saturated model in the sense that there are as many unknown elements involved in Λ as possible when we have two factors. However, as we proved in the previous paragraph, it is not an identified model because of nonexistence of unique linear transformation matrix. Though it is an unidentified model, it is worth mentioning the substantial meanings of the numbers of the rows and the numbers of the columns of the Λ. The numbers of the rows tell us how many residuals unique to the random coefficient β_j exist and the numbers of columns tell us how many common factors exist. In CFA, we decide the numbers of common factors mostly from the substantive knowledge, either theoretical knowledge or common sense.

As a more general than the model 1, and as an identifiable model, researchers may formulate the following model.

Model 3: Identified model, but a misspecified model

$$\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_1 & \lambda_4 \\ 0 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix}$$

This is an identified model and the model 1 is nested within this model, because if we fix $\lambda_3 = \lambda_4 = 0$, we obtain model 1. In order to test identifiability, we take two steps. That is, we first check the necessary conditions by (3-28). If the necessary conditions are satisfied, then we proceed to solve the covariance equations to see if indeed the model is identifiable.

Thus, we first check the three necessary conditions. The first one is satisfied because {the number of parameters in Tau} ($\frac{R(R+1)}{2}$) = 10 ($R = 4$) ≥ {the number of parameters estimated} (q) = 8 (4 + 3 + 1). It is clear that the second condition is also satisfied because there are 1's in each column of Λ at the position (1,1) and (3,2). To check the third condition, consider an invertible 2 by 2 matrix $T = \begin{pmatrix} x & y \\ u & v \end{pmatrix}$. Then,

$$\Lambda^* = \Lambda T = \begin{pmatrix} 1 & 0 \\ \lambda_1 & \lambda_4 \\ 0 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix} \begin{pmatrix} x & y \\ u & v \end{pmatrix} = \begin{pmatrix} x & y \\ \lambda_1 x + \lambda_4 u & \lambda_1 y + \lambda_4 v \\ u & v \\ \lambda_3 x + \lambda_2 u & \lambda_3 y + \lambda_2 v \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_1^* & \lambda_4^* \\ 0 & 1 \\ \lambda_3^* & \lambda_2^* \end{pmatrix}.$$

Thus we have conditions for Λ^* that need to be satisfied, which are $x = 1, y = 0, u = 0,$ and $v = 1$, that is, $T = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \mathbf{I}_2$. Therefore, the transformation is unique, so Λ is unique under linear transformation. Since all of the three necessary

conditions are satisfied, there is a hope that this is an identified model. In order to check whether it is an identified model, we see an algebraic solution. Simultaneous equations that we should solve for unknown parameters are obtained from $\tau = \Lambda\Psi\Lambda^T$. That is,

$$\begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{12} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{13} & \tau_{23} & \tau_{33} & \tau_{34} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_{44} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_1 & \lambda_4 \\ 0 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix} \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \lambda_1 & \lambda_4 \\ 0 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix}^T$$

$$= \begin{pmatrix} \psi_{11} & \psi_{12} \\ \lambda_1\psi_{11} + \lambda_4\psi_{12} & \lambda_1\psi_{12} + \lambda_4\psi_{22} \\ \psi_{12} & \psi_{22} \\ \lambda_3\psi_{11} + \lambda_2\psi_{12} & \lambda_3\psi_{12} + \lambda_2\psi_{22} \end{pmatrix} \begin{pmatrix} 1 & \lambda_1 & 0 & \lambda_3 \\ 0 & \lambda_4 & 1 & \lambda_2 \end{pmatrix}$$

$$= \begin{pmatrix} \psi_{11} & \lambda_1\psi_{11} + \lambda_4\psi_{12} & \psi_{12} & \lambda_3\psi_{11} + \lambda_2\psi_{12} \\ \lambda_1\psi_{11} + \lambda_4\psi_{12} & \lambda_1(\lambda_1\psi_{11} + \lambda_4\psi_{12}) + \lambda_4(\lambda_1\psi_{12} + \lambda_4\psi_{22}) & \lambda_1\psi_{12} + \lambda_4\psi_{22} & \lambda_3(\lambda_1\psi_{11} + \lambda_4\psi_{12}) + \lambda_2(\lambda_1\psi_{12} + \lambda_4\psi_{22}) \\ \psi_{12} & \lambda_1\psi_{12} + \lambda_4\psi_{22} & \psi_{22} & \lambda_3\psi_{12} + \lambda_2\psi_{22} \\ \lambda_3\psi_{11} + \lambda_2\psi_{12} & \lambda_3(\lambda_1\psi_{11} + \lambda_4\psi_{12}) + \lambda_4(\lambda_1\psi_{12} + \lambda_2\psi_{22}) & \lambda_3\psi_{12} + \lambda_2\psi_{22} & \lambda_3(\lambda_3\psi_{11} + \lambda_2\psi_{12}) + \lambda_2(\lambda_3\psi_{12} + \lambda_2\psi_{22}) \end{pmatrix}$$

Thus, we have simultaneous equations

$$\begin{cases} \psi_{11} = \tau_{11} \cdots\cdots (1) \\ \lambda_1\psi_{11} + \lambda_4\psi_{12} = \tau_{12} \cdots\cdots (2) \\ \psi_{12} = \tau_{13} \cdots\cdots (3) \\ \lambda_3\psi_{11} + \lambda_2\psi_{12} = \tau_{14} \cdots\cdots (4) \\ \lambda_1(\lambda_1\psi_{11} + \lambda_4\psi_{12}) + \lambda_4(\lambda_1\psi_{12} + \lambda_4\psi_{22}) = \tau_{22} \cdots\cdots (5) \\ \lambda_1\psi_{12} + \lambda_4\psi_{22} = \tau_{23} \cdots\cdots (6) \\ \lambda_3(\lambda_1\psi_{11} + \lambda_4\psi_{12}) + \lambda_2(\lambda_1\psi_{12} + \lambda_4\psi_{22}) = \tau_{24} \cdots\cdots (7) \\ \psi_{22} = \tau_{33} \cdots\cdots (8) \\ \lambda_3\psi_{12} + \lambda_2\psi_{22} = \tau_{34} \cdots\cdots (9) \\ \lambda_3(\lambda_3\psi_{11} + \lambda_2\psi_{12}) + \lambda_2(\lambda_3\psi_{12} + \lambda_2\psi_{22}) = \tau_{44} \cdots\cdots (10) \end{cases}$$

It can be seen that from Equations (2) and (6), we can solve for λ_1 and λ_4 because ψ_{11}, ψ_{12}, and ψ_{22} can be easily solved by Equations (1), (3), and (8) respectively. From Equations (4) and (9), we can solve for λ_2 and λ_3. Therefore, the model is identified. Actually it is an overidentified model because there are other ways to obtain the

solutions, For example, once we have solutions on $\psi_{11}, \psi_{12}, \psi_{22}, \lambda_1$, and λ_4, we can solve for λ_2 and λ_3 from Equations (7) and (10).

There are several characteristics that make this model one of the most substantively sensible models. First, the dimensionality was reduced from 4 to 2, and we specified that the number of unique factors is two. Specifying the first row of Λ as (1,0) and the third row as (0,1) make $u_{1j} = \eta_{1j}$, and $u_{2j} = \eta_{2j}$, which means that the unique randomness of math1 is the score of first factor, and the unique randomness of verbal1 is the score of second factor. The model defines that the variance factor 1 is exactly the variance of β_{1j}, the variability of math1 scores among examinee, and nothing else. The same thing can be said to the variance of β_{2j}, variability of verbal1 scores among examinees. This model is considered to be substantively most sensible.

Model 4: identified model, but a misspecified model

$$\Lambda = \begin{pmatrix} 1 \\ \lambda_1^* \\ \lambda_5^* \\ \lambda_3^* \end{pmatrix}, \ \Psi = \left(\psi_{11}^*\right).$$

This model is an identified model and is nested within model 1 so that we can apply the deviance test. The fact that this model is nested within the model 1 can be shown as follows. For model 4, we have

$$\Lambda\Psi\Lambda^T = \begin{pmatrix} 1 \\ \lambda_1^* \\ \lambda_5^* \\ \lambda_3^* \end{pmatrix} (\psi_{11}^*)(1 \quad \lambda_1^* \quad \lambda_5^* \quad \lambda_3^*)$$

$$= \begin{pmatrix} \psi_{11}^* & \lambda_1^*\psi_{11}^* & \lambda_5^*\psi_{11}^* & \lambda_3^*\psi_{11}^* \\ \lambda_1^*\psi_{11}^* & \lambda_1^{*2}\psi_{11}^* & \lambda_1^*\lambda_5^*\psi_{11}^* & \lambda_1^*\lambda_3^*\psi_{11}^* \\ \lambda_5^*\psi_{11}^* & \lambda_1^*\lambda_5^*\psi_{11}^* & \lambda_5^{*2}\psi_{11}^* & \lambda_5^*\lambda_3^*\psi_{11}^* \\ \lambda_3^*\psi_{11}^* & \lambda_1^*\lambda_3^*\psi_{11}^* & \lambda_5^*\lambda_3^*\psi_{11}^* & \lambda_3^{*2}\psi_{11}^* \end{pmatrix}.$$

We compare this expression to the model 1's counterpart:

$$\Lambda\Psi\Lambda^T = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{12} & \psi_{22} \end{pmatrix} \begin{pmatrix} 1 & \lambda_1 & 0 & 0 \\ 0 & 0 & 1 & \lambda_2 \end{pmatrix} = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \lambda_1\psi_{11} & \lambda_1\psi_{12} \\ \psi_{12} & \psi_{22} \\ \lambda_{12}\psi_{12} & \lambda_2\psi_{22} \end{pmatrix} \begin{pmatrix} 1 & \lambda_1 & 0 & 0 \\ 0 & 0 & 1 & \lambda_2 \end{pmatrix}$$

$$= \begin{pmatrix} \psi_{11} & \lambda_1\psi_{11} & \psi_{12} & \lambda_2\psi_{12} \\ \lambda_1\psi_{11} & \lambda_1^2\psi_{11} & \lambda_1\psi_{12} & \lambda_1\lambda_2\psi_{12} \\ \psi_{12} & \lambda_1\psi_{12} & \psi_{22} & \lambda_2\psi_{22} \\ \lambda_2\psi_{12} & \lambda_1\lambda_2\psi_{12} & \lambda_2\psi_{22} & \lambda_2^2\psi_{22} \end{pmatrix}.$$

In this expression, we let the correlation of latent variables be 1 to reflect our idea for model 4 that every 4 u_j's has completely correlated with each other. That is,

$\rho_{u_j u_j} = \dfrac{\psi_{12}}{\sqrt{\psi_{11}\psi_{22}}} = 1$, or, $\psi_{12} = \sqrt{\psi_{11}}\sqrt{\psi_{22}}$. Then let $\psi_{11} = \psi_{11}^*$, $\lambda_1 = \lambda_1^*$, $\psi_{22} = \lambda_5^*\sqrt{\psi_{11}^*}$, $\lambda_2 = \dfrac{\lambda_3^*}{\lambda_5^*}$. Then starting from model 4, we obtain model 1. Therefore, model 4 is nested within model 1. This process can be summarized by saying that after we set the correlation between two latent variables as 1, then we rescale the ψ_{22} and λ_2 by λ_5^*, where λ_5^* is newly specified as a element of the factor loading matrix Λ.

Chapter 4. Estimation and Inference

To estimate the parameters in the model (3-9), we adopt the method of full maximum likelihood (MLF) via Fisher scoring, adapting Longford's (1987) approach. We choose full maximum likelihood (MLF) as a method of estimation instead of restricted maximum likelihood (MLR)[1].

If we look at the model (3-9) from the general linear model point of view by letting

$$e_j = A_{2j}\theta_{2j} + r_j, \tag{4-1}$$

we have

$$Y_j = A_{1j}\theta_1 + e_j, e_j \sim N(0, V_j) \tag{4-2}$$

where

$$V_j = A_{2j}\Lambda\Psi\Lambda^T A_{2j}^T + \sigma^2 \mathbf{I}_{n_j}. \tag{4-3}$$

Let $\phi = (\phi_1, \phi_2, ..., \phi_Q)^T$ be a $Q \times 1$ vector of underlying parameters that produce V_j. It is well known that, conditional on the maximum likelihood estimate (MLE) of ϕ, the MLE of θ_1 is asymptotically orthogonal to the MLE of ϕ because the Hessian component

[1] The difference between MLR and MLF is that we specify a flat prior for the fixed effects in the case of MLR but not for MLF. The choice of model specification between MLF and MLR has pros and cons. The advantage of MLF is that the derivation is easier than MLR and is completely fit to the frequentist notion of Maximum Likelihood estimation. However, MLF underestimates variance component parameters because it doesn't take the uncertainty of the fixed effects. Thus, MLF produces biased estimates for variance components parameters whereas the corresponding MLR estimates are unbiased. However, MLF has advantage in terms of hypothesis testing because MLF allows deviance test for both fixed effects and random effects whereas MLR is only for random effects.

$E[\frac{\partial^2 l}{\partial\theta_1\partial\phi^T}]$ is 0, where l is the log likelihood of the model (3-9). The MLE of θ_1 is given by generalized least squares (GLS) as

$$\hat{\theta}_{1,MLE} = (A_1^T V^{-1} A_1)^{-1} A_1^T V^{-1} Y \tag{4-4}$$

where $A_1 = [A_{11}^T, A_{12}^T, \cdots, A_{1J}^T]^T$, a $N \times F$ matrix for $N = \sum_{j=1}^{J} n_j$, the total number of level-1 units, $V = \bigoplus_{j=1}^{J} V_j$, a $N \times N$ matrix, $Y = [Y_1^T, Y_2^T, \cdots, Y_J^T]^T$, a $N \times 1$ vector of observations. Note that $\oplus$ denotes the direct sum operator. And the asymptotic covariance matrix of $\hat{\theta}_{1,MLE}$ is

$$D[\hat{\theta}_{1,MLE}] = (A_1^T V^{-1} A_1)^{-1} \tag{4-5}$$

where V is evaluated at $\hat{\phi}_{MLE}$. Since $E[\frac{\partial^2 l}{\partial\theta_1\partial\phi^T}]$=0[2], we just need a series of scoring iterations with respect to ϕ to obtain $\hat{\phi}_{MLE}$. After we obtain $\hat{\phi}_{MLE}$, we compute $\hat{\theta}_{1,MLE}$ by formula (4-4).

The log likelihood based on Equation (4-2) is

$$l(\theta_1, \phi; Y) \equiv \log L(\theta_1, \phi; Y) = -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|V| - \frac{1}{2}e^T V^{-1} e, \tag{4-6}$$

where $e = \bigoplus_{j=1}^{J} e_j$. Since Y is a random sample of size J of Y_j,

[2] This is generally true when we take derivative of l twice with respect to the fixed effects θ_1 and the random effects ϕ successively, and then take the expectation with respect to the distribution.

$$l(\theta_1, \phi, ; Y) = \sum_{j=1}^{J} l_j(\theta_1, \phi ; Y_j) \tag{4-7}$$

where

$$l_j(\theta_1, \phi ; Y) \equiv \log L_j(\theta_1, \phi ; Y_j) = -\frac{n_j}{2}\log(2\pi) - \frac{1}{2}\log|V_j| - \frac{1}{2} e_j^T V_j^{-1} e_j . \tag{4-8}$$

Longford (1987) gives the score vector and Hessian of the log-likelihood function (4-8) in terms of element wise formula. Raudenbush (1994) showed the formulae in more compact matrix form. The key results are for the $q \times 1$ score vector

$$S_j \equiv \frac{\partial l_j}{\partial \phi} = \frac{1}{2}\left(\frac{\partial\, vec V_j}{\partial \phi}\right)^T (vec M_j) \tag{4-9}$$

where

$$M_j = V_j^{-1}(e_j e_j^T - V_j)V_j^{-1} \tag{4-10}$$

and

$$S = \sum_{j=1}^{J} S_j . \tag{4-11}$$

For the $q \times q$ Hessian matrix,

$$H_j \equiv E\left[\frac{\partial^2 l_j}{\partial \phi \partial \phi^T}\right] = -\frac{1}{2}\left(\frac{\partial\, vec V_j}{\partial \phi}\right)^T (V_j^{-1} \otimes V_j^{-1})\left(\frac{\partial vec V_j}{\partial \phi}\right) \tag{4-12}$$

and

$$H = \sum_{j=1}^{J} H_j . \tag{4-13}$$

We apply the above general formulae to obtain $\hat{\phi}_{MLE}$. The details of the derivation and the algorithm are in the Appendix.

To implement the Fisher scoring algorithm, we need a starting value of the parameters in the variance-covariance matrix V_j. Though it is arbitrary, we need to provide a good starting value to obtain the MLE within the parameter space and its quicker convergence. Though V_j is a function of three components, i.e., λ, ψ, and σ^2 (See Equation (A-4) in the Appendix), the hard part of providing a good starting value is in λ and ψ, which are the vector of unknown elements of Λ and unique elements of Ψ (Starting values for the estimate of σ^2 can be found in the section of Starting Values in the Appendix.). One way to obtain the starting values for the estimates of λ and ψ is by analogy to the factor analysis. That is, we execute principal component factor analysis on the positive semi-definite $\bar{\tau}$, which is obtained by using the same method as current HLM2, and then take advantage of the pre-specified structure of Λ to find the initial estimate of Λ, which is denoted by $\hat{\Lambda}^{(0)}$. Then solving $\tau = \Lambda\Psi\Lambda^T$ for Ψ and substituting Λ by $\hat{\Lambda}^{(0)}$ and τ by $\bar{\tau}$, we can obtain $\hat{\Psi}^{(0)}$, the initial estimate of Ψ. The details of the step was shown in Appendix.

Finally I note that in order to obtain approximate standard errors for $\hat{\phi}_{MLE}$, we will compute the information matrix by

$$Info. = -H \tag{4-14}$$

where H is evaluated at $\hat{\phi}_{MLE}$. Then, the asymptotic standard errors for $\hat{\phi}_{MLE}$ are computed by

$$s.e.(\hat{\phi}_r) = \frac{1}{\sqrt{Info._{rr}}} (r = 1,..., Q) \tag{4-15}$$

where ϕ_r is the rth element of vector ϕ and $Info._{rr}$ is the corresponding rth diagonal element of the information matrix (*Info.*). For the standard error of $\hat{\theta}_1$, the MLE of fixed effects θ_1, we will compute the asymptotic covariance

$$D[\hat{\theta}_1] \approx (A_1^T V^{-1} A_1)^{-1} \tag{4-16}$$

where V is evaluated at $\hat{\phi}$, the MLE of covariance structure parameter ϕ. Then the standard error of the rth element of a vector $\hat{\theta}_1$ is given by

$$s.e.(\hat{\theta}_{1r}) = [(A_1^T V^{-1} A_1)^{-1}]_{rr} \tag{4-17}$$

for r = 1,2,..., F. Note that since $\hat{\phi}_r$ and $\hat{\theta}_{1r}$ are known to be asymptotically normally distributed, we can use them for the inference by keeping in mind that it's an approximation.[3]

[3] The large sample normal approximation for the standard error of $\hat{\phi}$ is poor in most applications so that other techniques are required in practice (Bryk & Raudenbush, 1992, Chapter 3).

Chapter 5. Computational Examples

In this chapter, I will demonstrate how the proposed model can be applied to real data sets and when it is useful to do so. I first show that the HLM2 factor analysis model (HLM2F) includes the standard HLM2 model as a submodel, by using a model formulated in Bryk & Raudenbush (1992) for the High School and Beyond data. The next example shows a case when the HLM2F model is useful, using the data from Huttenlocher et al. (1991) on children's vocabulary growth. The third example shows how to specify the factor analysis model when the Tau-matrix (the level-2 variance-covariance matrix) is relatively large such that specification of the factor structure is less obvious. The fourth example uses artificial data with known parameters. The purpose of demonstrating this example is not only to provide a check for the validity of the computation, but also to demonstrate how to formulate a reasonable model, i.e., a substantively meaningful and an identifiable model. Further, it illustrates how to evaluate and correct mis-specified models.

5-1. **High School and Beyond**

I use the High School and Beyond (HS & B) to demonstrate that the standard HLM2 model is a submodel of the two-level Hierarchical Linear Model with Factor structure (HLM2F).

The data is a subsample from the 1982 High School and Beyond Survey, and includes information on 7,185 students nested within 160 schools. In Table 4.5 on page 72 of Bryk & Raudenbush (1992), the results of the following model are presented:

HLM2 model:

L-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(SES - \overline{SES_{.j}})_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

L-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Sector)_j + \gamma_{02}(\overline{SES_{.j}})_j + u_{0j} \qquad (5\text{-}1)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Sector)_j + \gamma_{12}(\overline{SES_{.j}})_j + u_{1j},$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{pmatrix}\right).$$

Here, Y_{ij} is the math achievement score for student i in school j; SES_{ij} is a measure of the socio-economic status (SES) of student i in school j; $\overline{SES_{.j}}$ is the sample mean of SES of the school j; and $Sector_j$ is the indicator variable taking on a value of 'one' for Catholic schools and 'zero' for public schools.

Since the model which I developed in Chapter 3 uses the full maximum likelihood with the Fisher scoring algorithm and the default HLM2 uses the restricted maximum likelihood with the EM algorithm, we set the estimation method of HLM2 as comparable as possible to the HLM2F. This can be done by setting the HLM2 optional specification to MLF (Full maximum likelihood) and the number of Fisher iterations to 1. Setting the number of Fisher iterations to 1 specifies the program to run all the way by Fisher

scoring, but if the algorithm fails for some reason[1], then the algorithm switches back to the EM algorithm.

The results of HLM2 by MLF with the number of Fisher scoring = 1 are in the second column of Table 5.1.

The HLM2F model equivalent to the above model can be specified by setting the factor loading matrix as the identity matrix of the size of the number of random effects in HLM2. This can be shown as follows.

HLM2F model:

L-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}(SES - \overline{SES_{\cdot j}})_{ij} + \varepsilon_{ij}, \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$ (Same as L-1 model of Equation (5.1))

L-2:

$$\beta_{0j} = \gamma_{00} + \gamma_{01}(Sector)_j + \gamma_{02}(\overline{SES_{\cdot j}})_j + 1 \cdot \eta_{1j} + 0 \cdot \eta_{2j} \qquad (5\text{-}2)$$

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(Sector)_j + \gamma_{12}(\overline{SES_{\cdot j}})_j + 0 \cdot \eta_{1j} + 1 \cdot \eta_{2j}$$

where

$$\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}\right).$$

In matrix form, the level-2 model can be written as

$$\beta_j = W_j\gamma + \Lambda\eta_j \qquad (5\text{-}3)$$

[1] Most typical reason of this failure is that Fisher will produce negative definite $\hat{\tau}$.

where $W_j = \begin{pmatrix} 1 & Sector_j & \overline{SES_{\cdot j}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & Sector_j & \overline{SES_{\cdot j}} \end{pmatrix}$,

$\gamma = \left(\gamma_{00},\ \gamma_{01},\ \gamma_{02},\ \gamma_{10},\ \gamma_{11},\ \gamma_{12}\right)^T$, $\Lambda = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$, and $\eta_j = \left(\eta_{0j},\ \eta_{1j}\right)^T$.

Thus if we set $\Lambda = \mathbf{I}_2$ in the HLM2 model, then $u_j = \eta_j$ and $\tau = \Lambda\Psi\Lambda^T = \Psi$.

The results using HLM2F are in the third column of Table 5.1. As can be seen by comparing the results, the HLM2F reproduced almost exactly the same estimates for all of the parameters including the standard errors. This is evidence that the newly developed program is working.

Table 5.1 Results of the Analysis for High School and Beyond Survey

	HLM2	HLM2F
# iterations until convergence	5	3
Fixed effects estimates		
γ_{00}	12.128 (0.197)[2]	12.128(0.197)
γ_{01}	1.227 (0.303)	1.227(0.303)
γ_{02}	5.332 (0.366)	5.332(0.366)
γ_{10}	2.946 (0.154)	2.946(0.154)
γ_{11}	-1.644 (0.237)	-1.644(0.237)
γ_{12}	1.042 (0.296)	1.042(0.296)
Random effects estimates		
σ^2	36.721 (0.626)	36.721(0.619)
τ_{00} (or ψ_{00} for HLM2F)	2.317 (0.355)	2.317(0.353)
τ_{10} (or ψ_{10} for HLM2F)	0.188 (0.196) (0.483 as corr.)	0.188(0.193)
τ_{11} (or ψ_{11} for HLM2F)	0.065 (0.208)	0.065(0.204)
Log-likelihood at convergence	-23247.30	-23248.22
# parameters estimated	10	10
Deviance	46494.592	46496.44

[2] Note. ( ) represents the standard error computed from the Information matrix.

5-2. **Infant Vocabulary Growth**

The purpose of demonstrating the analysis of infant vocabulary growth data is to show where it is useful to apply the HLM2F model. The data come from a recent study of children's vocabulary development during the second year of life ((Huttenlocher, Haight, Bryk, Seltzer, & Lyons, 1991). Huttenlocher et al. investigated the relationship between a child's early vocabulary acquisition and maternal use of language, hypothesizing that exposure to the mother's spoken language has a positive relationship to the growth of the vocabulary of the young child. Gender differences in vocabulary growth were also considered (Huttenlocher et al., 1991).

Two groups of mother-infant pairs, with six boys and five girls in each, provided from 5 to 7 observations per infant in their early stages of life. More specifically, the first group consisted of 11 children who were observed at their home on six or seven occasions at 2-month intervals during the period from 14 to 26 months of age. For some cases, the 14-months data point was missing. The second group consisted of another 11 children who were observed at 16, 20, and 24 months. The time dimension was $(Age - 12)_{ij}$ months, assuming that at 12 months a child's vocabulary size was zero.

Huttenlocher et al. (1991) first formulated the following full quadratic unconditional model.

L-1: $$Y_{ij} = \beta_{0j} + \beta_{1j}(Age - 12)_{ij} + \beta_{2j}(Age - 12)^2_{ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

L-2:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{5-4}$$

$$\beta_{1j} = \gamma_{10} + u_{1j}$$

$$\beta_{2j} = \gamma_{20} + u_{2j},$$

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{01} & \tau_{11} & \tau_{12} \\ \tau_{02} & \tau_{12} & \tau_{13} \end{pmatrix} \right)$$

Obtaining the results that H_0: $\tau_{00} = 0$ and $\gamma_{00} = 0$ and $\gamma_{10} = 0$ can hold statistically and knowing that $\hat{Corr}(u_{1j}, u_{2j}) \approx 1$, they formulated the following final level-1 model which left only the quadratic term:

L-1:

$$Y_{ij} = \beta_{2j}(Age - 12)_{ij}^2 + \varepsilon_{ij},\ \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

At level-2, they modeled the rate of acceleration by the individual characteristics such as group membership, gender, and the amount that the mother spoke to the child.

L-2:

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(Group)_j + \gamma_{22}(Gender)_j + \gamma_{23}\log(Momspeak)_j + u_{2j}, \tag{5-5}$$

$$(u_{2j}) \overset{iid}{\sim} N((0),(\tau_{00})).$$

where $(Group)_j = 1$ if the child is in group 1, and 0 otherwise; $(Gender)_j = 1$ if the child is a girl, and 0 if a boy; and $\log(Momspeak)_j$ is the natural logarithm of the number of

words that the mother spoke to the child j, measured once at the first occasion of the observation. The results for this model are in the second column of Table 5.2[3].

The fact that the estimate of the correlation between u_{1j} and u_{2j} was high may not necessarily indicate that the rate of growth does not vary from child to child, which is a question that was left in Huttenlocher et al. (1991)'s model. To investigate this point, Bryk & Raudenbush (1992) used the following model:

L-1:

$$Y_{ij} = \beta_{1j}(Age-12)_{ij} + \beta_{2j}(Age-12)^2_{ij} + \varepsilon_{ij},\ \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$$

L-2:

$$\beta_{1j} = u_{1j}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(group)_j + \gamma_{22}(gender)_j + \gamma_{23}\log(Momspeak)_j + u_{2j} \qquad (5\text{-}6)$$

where

$$\begin{pmatrix} u_{1j} \\ u_{2j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{12} & \tau_{22} \end{pmatrix}\right).$$

The results showed that $\hat{\tau}_{11} = 16.94$ (See the row for τ_{11} at column 3 in Table 5.2), which was significantly different from 0, and the u_{1j} and u_{2j} were highly correlated. This suggested that it was not necessary to estimate the covariance τ_{12}, because the null

[3] The results were obtained by reanalyzing the model by HLM2 version 4.92, specifying full maximum likelihood (MLF) and # Fisher acceleration = 1. This specification induced slightly different results from Huttenlocher et al. (1991) which used the restricted maximum likelihood (MLR) and EM algorithm. The same thing can be said in the next results of Bryk and Raudenbush's model.

hypothesis $\tau_{12} = \sqrt{\tau_{11} \cdot \tau_{22}}$ held, in the population, at .05 level, but it was necessary to estimate both variances, τ_{11} and τ_{22} because those were not zero.

The above result implies a factor analytic type model. Reparameterizing the τ_{22} as $\tau_{22} = \lambda_{21}^2 \tau_{11}$ by using the variance ratio $\lambda_{21}^2 = \frac{\tau_{22}}{\tau_{11}}$, and then setting $\tau_{12} = \lambda_{21}\tau_{11}$ gives the same constraint as $\tau_{12} = \sqrt{\tau_{11} \cdot \tau_{22}}$. Thus, we obtain a factor analytic type (HLM2F) model by giving certain constraints on the elements of τ matrix of the original HLM2 model, which in turn, implies that the HLM2F model that will be formulated is nested within the HLM2 model.

This idea is formally formulated by a factor analysis type model. That is,

In L-2 model, we let

$$u_{0j} = \eta_j$$
$$u_{1j} = \lambda_{21}\eta_j,$$

or, in matrix form,

$$u_j = \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} = \begin{pmatrix} 1 \\ \lambda_{21} \end{pmatrix} \begin{pmatrix} \eta_j \end{pmatrix} = \Lambda \eta_j, \tag{5-7}$$

and the covariance equation is

$$D[u_j] = \Lambda \Psi \Lambda^T = \begin{pmatrix} \psi_{11} & \lambda_{21}\psi_{11} \\ \lambda_{21}\psi_{11} & \lambda_{21}^2\psi_{11} \end{pmatrix}. \tag{5-8}$$

Thus, we formulate the following model:

L-1:

$$Y_{ij} = \beta_j (Age - 12)_{ij} + \beta_{2j} (Age - 12)_{ij}^2 + \varepsilon_{ij} \qquad \varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$$

L-2:

$$\beta_{1j} = \eta_{1j} \tag{5-9}$$

$$\beta_{2j} = \gamma_{20} + \gamma_{21}(Group)_j + \gamma_{22}(Gender)_j + \gamma_{23}\log(MomSpeak)_j + \lambda_{21}\eta_{1j}$$

where $\eta_{1j} \overset{iid}{\sim} N(0, \psi_{00})$.

The results of this reduced model are in the fifth column of Table 5.2. In order to use the likelihood ratio test to examine whether the above factor model is statistically acceptable, we run the standard HLM2 model in two ways, one is to run the HLM2 software and another is to run the HLM2F program by setting $\Lambda = \mathbf{I}_2$. Those results are in the third column of the table below, under the caption HLM2 (Bryk and Raudenbush's model), and in the fourth column, under the caption HLM2F (Bryk and Raudenbush's model)[4].

[4] It should be noted that the results for the standard HLM2 model executed by HLM2F showed a close match to the HLM2 results. However, a run by HLM2F was problematic because the estimate of Ψ became a negative definite matrix in the early stage of scoring iteration even though several different starting values were tried. The values on the table were obtained when we ignored the fact that the estimate of Ψ became a negative definite matrix during the iteration. For this reason, we take the results by HLM2F for the saturated model as a reference. On the other hand, HLM2 was executed by choosing MLF and the number of Fisher iterations = 1. This specification lets HLM2 run repeatedly under Fisher scoring algorithm as long as the estimate of Ψ does not become a negative definite matrix during the iteration; if it does, then HLM2 returns to the EM algorithm, which is a default algorithm. Thus, probably this is how HLM2 produced the results: HLM2 first tried to estimate by scoring but it failed, and then used EM.

Table 5.2 Results of the Analysis for the Infant Vocabulary Growth Data[5]

	HLM2 (Huttenlocher et al's model)	HLM2 (Bryk and Raudenbush's Model)	HLM2F (Bryk and Raudenbush's Model)	HLM2F (Reduced Model)
# iterations until convergence	6	8	4	8
Fixed effects estimates				
γ_{20}	2.150(0.277)	2.088(0.140)	2.181(0.139)	2.180(0.139)
γ_{21}	0.802(0.345)	0.599(0.290)	0.593(0.288)	0.598(0.289)
γ_{22}	-1.105(0.327)	-0.904(0.279)	-0.902(0.278)	-0.904(0.279)
γ_{23}	0.886(0.327)	0.827(0.268)	0.826(0.267)	0.827(0.268)
Random effects estimates				
σ^2	819.366(113.622)	707.866(107.832)	711.025(98.197)	708.412(98.233)
τ_{11}(ψ_{11} for HLM2F)	N.A.	16.940(15.402)	16.3266(13.438)	16.8429(5.256)
τ_{21}	N.A.	1.709(0.978) (0.985 as corr.)	1.77651(0.864)	N.A.
τ_{22}	0.566(0.177)	0.178(0.128)	0.172(0.119)	N.A.
λ_{21}	N.A.	N.A.	N.A.	0.102(0.0262)
Log-likelihood at convergence	-639.22	-632.505	-632.502	-632.504
# parameters estimated	6	8	8	7
Deviance	1278.436	1265.010	1265.004	1265.008

We compare the results in column 3 for the full model (Bryk and Raduenbush's model) and those in column 5 for the reduced model (reduced by factor model). First, notice that the point estimates and their estimates of the standard error for the fixed effects are almost the same for the two models. The estimates of level-1 error variance (σ^2) are also almost the

[5] Note:
1. HLM2 results were obtained by setting Full maximum likelihood and # Fisher acceleration =1. When we set the # Fisher acceleration =5, then the results in column 3 took 867 iterations to converge.
2. ( ) is the standard error.
3. The cutoff criterion for getting out of the Fisher scoring loop for HLM2F is 'relative change in the squared length of parameter estimates' < 10^{-9}, which is the same value as HLM2, though HLM2 uses changes in log-likelihood.

(σ^2) are also almost the same, but there is a small difference for the standard errors. For the level-2 variance-covariances, we can recover the estimates of the original variance-covariances if we compute them using the points estimates in the reduced model and the covariance equation in Equation (5-8) for the original Tau-matrix. Thus:

$\hat{\lambda}_{21}\hat{\psi}_{11} = 0.10229 \times 16.8429 = 1.72286,$

$\hat{\lambda}_{21}^2\hat{\psi}_{11} = 0.10229^2 \times 16.8429 = 0.176231.$

These values closely match those in the second column of Table 5.2, i.e., $\hat{\tau}_{12} = 1.709$ and $\hat{\tau}_{22} = 0.178$ respectively, which are the corresponding elements in the τ matrix. This means that we can recover the original τ estimate, and this fact is especially useful when we cannot directly obtain the estimate of τ, which is often the case when $\hat{\tau}$ is not full rank, as happened in this data when we used the Fisher scoring in HLM2F.

The comparison of the deviances for the nested models can be used to assess model fit using a likelihood ratio test. The deviances for the full model (1265.010) and for the reduced model (1265.008) are basically the same. Thus, the deviance test clearly shows that the reduced model is statistically acceptable.

5-3. **Scholastic Aptitude Test (SAT) Meta-analysis of Coaching Effects**

The purpose of this demonstration is to give an example where the τ matrix is of relatively large dimensions. When the size of τ is large, we can formulate many different factor structures, which adds a complexity to the modeling. At this point, discipline specific theory and knowledge about particular data and variables should inform decisions regarding the number of factors (M), and which elements of factor loading matrix (Λ) to fix and what values to use, etc. After deciding on the general framework of the factor structure, mathematical knowledge of model identification can be applied to decide whether the model in mind is an identified model.

In the SAT meta-analysis for coaching effect data, we create a 4×4 τ matrix. The data consist of 48 studies on coaching effects on the SAT scores (Becker, 1990). Only 46 studies were analyzed for our analysis because two studies had no information on coaching hours. The SAT scores were reported for the two subtests, SAT-Math (SAT-M) and SAT-Verbal (SAT-V). Each study provides a part of the information on standardized mean change scores between pretest and posttest on SAT-M, SAT-V for coached and uncoached groups. A study is considered to be complete if it provides four standardized mean change scores, i.e., standardized mean change score for the coached group on SAT-M, standardized mean change score for the uncoached group on SAT-M, standardized mean change score for the coached group on SAT-V, and standardized mean change score for the uncoached group on SAT-V. Table 5.3 shows a number of studies classified by available standardized mean change scores and whether a control group

(uncoached group) was used in the study. From that table, we see that only 19 studies, which is about 42% of the studies, had a complete set of information.

Table 5.3 Classification of SAT Studies

Existence of Control Group

Available Standardized Mean Change	Yes	No	Row Subtotal
Both SAT-M and SAT-V	19	2	21
SAT-M only	1	3	4
SAT-V only	13	8	21
Column Subtotal	33	13	Total 46

One of the strengths of a hierarchical analysis is that diverse patterns of data information can be put together and all the information can be used for statistical inference assuming that data are missing at random (Little and Rubin, 1989). This will be the method of analysis used on the above meta-analysis data, which utilizes data from 27 studies (about 58 %) that are incomplete. Unless using a hierarchical model, information from more than half of the studies can be totally discarded or it will be used less effectively. In a small data set such as this meta-analysis data, it is too costly.

Model:

Let d_{ij} be the standardized mean change score in the ith observation of the jth study, which can be defined as

$$d_{ij} = \frac{4(n_{ij}-2)}{4n_{ij}-5}\left(\frac{\overline{Y}_{2ij} - \overline{Y}_{1ij}}{S_{Y_{2ij}}}\right) \qquad (5\text{-}10)$$

where $\overline{Y}_{2ij}$ is the sample post-test average for observation i of study j; $\overline{Y}_{1ij}$ is the sample pre-test average for observation i of study j, $S_{Y_{2ij}}$ is the sample standard deviation of the post test for observation i of study j which is assumed to well approximates the pooled sample standard deviation, and the coefficient $\frac{4(n_{ij}-2)}{4n_{ij}-5}$ for $\frac{\overline{Y}_{2ij} - \overline{Y}_{1ij}}{S_{Y_{2ij}}}$ is used as a multiplying factor to ensure that d_{ij} is an unbiased estimator of the population standardized mean change $\delta_{ij} = \frac{\mu_{2ij} - \mu_{1ij}}{\sigma}$ (Becker, 1988), where μ_{2ij} is the population post-test mean for observation i of study j, μ_{1ij} is the population pre-test mean for observation i of study j, and σ is the population standard deviation that is common to both pre-test and post-test.

Let X_1 be the indicator for SAT-M score for the coached group, X_2 be the indicator for SAT-M score for the uncoached group, X_3 be the indicator for SAT-V score for the coached group, and X_4 be the indicator for SAT-V score for the uncoached group. Thus, X_{1ij} = 1 if d_{ij} is the observation of SAT-M score for coached group, and 0 otherwise; X_{2ij} = 1 if d_{ij} is the observation of SAT-M score for uncoached group, and 0 otherwise, X_{3ij} = 1 if d_{ij} is the observation of SAT-V score for coached group, and 0 otherwise; X_{4ij} = 1 if d_{ij} is the observation of SAT-V score for uncoached group, and 0 otherwise. Defining the variables as above, we formulate the level-1 HLM model:

L-1:

$$d_{ij} = \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \varepsilon_{ij}, \qquad (5\text{-}11)$$

where $i = 1,...p_j$ for p_j is the number of observations of standard mean difference d_{ij} available for the study j, and $j = 1,2, ...,46$. Within the study, the error ε_{ij}'s are correlated because even within a study, sources of errors come from the same set of people. This is the difference between this model and the standard HLM model where the level-1 errors are assumed independent. The dependency of the within-study errors occurred because the model is multivariate, which implies that the data have common sources of variation within a study. The variances and covariances of the covariance matrix of the vector of within-study error $\varepsilon_j = (\varepsilon_{1j},...,\varepsilon_{p_j j})^T$ ($V_j = D[\varepsilon_j]$, a $p_j \times p_j$ symmetric matrix) can be computed by the following formula (Becker, 1990):

$$Var(\varepsilon_{ij}) \equiv V_{ij} = \frac{2(1-\rho_{PP})}{n_j} + \frac{\delta_{ij}^2}{2n_j}, \qquad (5\text{-}12)$$

where n_j is the number of subjects in the study j, and ρ_{PP} is the population pretest-posttest correlation. In this analysis, the value of $\rho_{PP} = 0.88$ will be used as the approximate population correlation between pre- and posttest SAT scores for both SAT-M and SAT-V for all subjects, following DerSimonian and Laird (1983). Similarly the covariances of the covariance matrix of the vector of within-study error can be computed by the following formula (Becker, 1990).

$$Cov(\varepsilon_{ij}, \varepsilon_{i'j}) \equiv V_{ii'j} = \frac{\rho_{ii'j}}{n_j} + \frac{\delta_{ij}\delta_{i'j}\rho_{ii'j}^2}{2n_j}, \qquad (5\text{-}13)$$

where $\rho_{ii'j}$ is the population correlation between the dependent variables associated with estimated standardized mean differences i and i'. This correlation may be estimated from the sample, deduced from published test information or imputed on the basis of past research. In the SAT meta-analysis case, we use $\rho_{ii'j} = 0.66$ if the pair of (i , i') is Math and Verbal in the same experimental group (either in the coached group or uncoached group), which is taken from the test manual (Gleser & Olkin, 1994). If the pair of the observation unit (i , i') is in the different groups, i.e., coached and uncoahed group, $\rho_{ii'j} = 0.00$ since those observations are independent.

As we saw in the above, the within-study error variance-covariance are known in the meta-analysis settings. In HLM terms, this type of model is said to be the V-known model.

The level-2 model is a multivariate regression model and the time for coaching in hours is used to model β_{1j} and β_{3j}, the means of standardized mean difference for SAT-M and SAT-V for the coached groups.

L-2:

$$\begin{aligned}
\beta_{1j} &= \gamma_{10} + \gamma_{21}(\textit{Coaching Hours})_j + u_{1j} \\
\beta_{2j} &= \gamma_{20} + u_{2j} \\
\beta_{3j} &= \gamma_{20} + \gamma_{21}(\textit{Coaching Hours})_j + u_{3j} \\
\beta_{4j} &= \gamma_{40} + u_{4j}
\end{aligned} \qquad (5\text{-}13)$$

where

$$\begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{21} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{31} & \tau_{32} & \tau_{33} & \tau_{34} \\ \tau_{41} & \tau_{42} & \tau_{43} & \tau_{44} \end{pmatrix} \right),$$

and $(Coaching\ Hours)_j$ is the hours of coaching for the study j.

This model is substantively interesting and meaningful by two reasons: One is that if we center *Coaching Hours* around zero as we just did in the Equation (5-13), then the fixed effect γ_{10} is the expected standardized mean change for SAT-M for the coached groups at the absence of coaching and thus $\gamma_{10} - \gamma_{20}$ represents the expected difference between coached and uncoached groups for SAT-M under no treatment (coaching), which is a possible result because some of the studies were not randomized experiments. Similar interpretation can be made for $\gamma_{30} - \gamma_{40}$, which is the expected difference between coached and uncoached groups for SAT-V under no treatment. The second reason is that If we center *Coaching Hours* around its grand mean, then $\gamma_{10} - \gamma_{20}$ represents the mean gains in the experimental groups compared to the standardized mean change of the control groups for SAT-M, and $\gamma_{30} - \gamma_{40}$ is the mean gains in the experimental groups compared to the standardized mean change of the control groups for SAT-V, assuming that the model is correctly specified[6].

To solve V-known HLM model, we capitalize the fact that the level-1 error variance-covariance matrix V_j is known for all j; j = 1,2,...,J, where J is the number of

[6] Actually model misspecification is possible. Kalaian and Raudenbush (1996) used the natural logarithmic transformation of the hours of coaching, being concerned with a curvilinear relationship from the observations of the scatterplots. Thus, including quadratic term may show a better fit. Here I used a linear model for simplicity of demonstrating the point.

studies that was used in Kalaian and Raudenbush (1996). That is, using Cholesky factorization, $V_j = F_j F_j^T$, where F_j is a $p_j \times p_j$ lower triangular positive definite matrix, we transform the level-1 within study model in Equation (5-11) by multiplying F_j^{-1} from the left. Then the transformed level-1 model has i.i.d. error variance for all the studies with the fixed unit variance. Therefore, we can apply HLM2 and HLM2F software by fixing the level-1 error variance (σ^2) = 1.

The results of this model by HLM2 is shown in the second column of Table 5.4. The results for the same model by HLM2F was not obtained because the estimate of Ψ becomes negative definite during the early stage of scoring iteration. The estimate of τ was

$$\hat{\tau} = \begin{pmatrix} 0.0775 & 0.0874 & -0.0360 & -0.0156 \\ 0.0874 & 0.1045 & -0.0352 & -0.0191 \\ -0.0360 & -0.0352 & 0.0532 & 0.0364 \\ -0.0156 & -0.0191 & 0.0364 & 0.0334 \end{pmatrix}, \tag{5-14}$$

and as a correlation,

$$\hat{\rho}_\tau = \begin{pmatrix} 1.000 & 0.972 & -0.562 & -0.307 \\ 0.972 & 1.000 & -0.472 & -0.324 \\ -0.562 & -0.472 & 1.000 & 0.863 \\ -0.307 & -0.324 & 0.863 & 1.000 \end{pmatrix}. \tag{5-15}$$

As we can see from the correlation estimates, $\hat{\tau}$ is almost singular and has two high correlation blocks, which caused the slow convergence that is shown by the very large number of iterations in HLM2 as 2911 EM iterations, and caused the inability of obtaining the estimation by HLM2F which utilized Fisher scoring.

Table 5.4 Results of the Analysis for the SAT Coaching Effects Data[7]

	HLM2 (Full model)	HLM2F (Reduced model a)	HLM2F (Reduced model b)
# iterations until convergence	2911	13	8
Fixed effects estimates			
γ_{10}	0.290(0.071)	0.317(0.0692)	0.284(0.0679)
γ_{11}	0.00716(0.000896)	0.00701(0.00110)	0.00760(0.00115)
γ_{20}	0.263(0.0709)	0.254(0.0614)	0.253(0.0655)
γ_{30}	0.183(0.0517)	0.176(0.0459)	0.187(0.0537)
γ_{31}	0.00467(0.00158)	0.00481(0.00091)	0.00455(0.00123)
γ_{40}	0.140(0.0357)	0.1409(0.0321)	0.1417(0.0343)
Random effects estimates			
σ^2	N.A.	N.A.	N.A.
τ_{11} (ψ_{11} for HLM2F a & b)	0.0775(0.0241)	0.0867(0.0243)	0.0812(0.0208)
τ_{21}	0.0874(0.0269)	N.A.	N.A.
τ_{22}	0.104(0.0330)	N.A.	N.A.
τ_{31} (ψ_{21} for HLM2F a & b)	-0.0360(0.0154)	-0.0269(0.0135)	-0.0325(0.00893)
τ_{32}	-0.0352(0.0180)	N.A.	N.A.
τ_{33} (ψ_{22} for HLM2F a & b)	0.0532(0.0137)	0.0496(0.0117)	0.0536(0.00811)
τ_{41} (ψ_{31} for HLM2F b)	-0.0156(0.0125)	N.A.	-0.0167(0.00673)
τ_{42}	-0.0191(0.0146)	N.A.	N.A.
τ_{43} (ψ_{32} for HLM2F b)	0.0364(0.0102)	N.A.	0.0329(0.00325)
τ_{44} (ψ_{33} for HLM2F b)	0.0334(0.00958)	N.A.	0.0324(0.00464)
λ_{21}	N.A.	1.010(0.0541)	1.122(0.0717)
λ_{42}	N.A.	0.817(0.0556)	N.A.
Log-likelihood at convergence	-279.689	-307.693	-284.424
# parameters estimated	16	11	13
Deviance	559.378	615.386	568.848

[7] Note
1. ( ) is the standard error
2. The cutoff criterion for getting out of the Fisher scoring loop for HLM2F is 'relative change in the squared length of parameter estimates' < 10^{-6}, which is the same value as HLM2, though HLM2 uses changes in log-likelihood.

From the pattern of $\hat{\tau}$, we can imagine a simple pattern of factor structure. That is, we formulate the factor model for level-2 error vector $u_j = \left(u_{1j},\ u_{2j},\ u_{3j},\ u_{4j}\right)^T$ as

$$\begin{aligned} u_{1j} &= \eta_{1j}, \\ u_{2j} &= \lambda_{21}\eta_{1j}, \\ u_{3j} &= \eta_{2j}, \\ u_{4j} &= \lambda_{42}\eta_{2j}. \end{aligned} \tag{5-16}$$

or, in matrix form,

$$u_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} = \Lambda\eta_j, \tag{5-17}$$

where $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix}$, and $\eta_j = \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix}$.

The pattern of factor loading matrix Λ and the length of factor score vector η_j characterizes our idea about how the level-2 errors are correlated. That is, we hypothesize that the 4 level-2 random errors ($u_j = \left(u_{1j},\ u_{2j},\ u_{3j},\ u_{4j}\right)^T$) after controlling for the coaching time can be represented by two unique level-2 unit latent random errors ($\eta_j = \left(\eta_{1j},\ \eta_{2j}\right)^T$), and the SAT-M unique variability for coached and uncoahed groups ($\left(u_{1j},\ u_{2j}\right)$) can be explained by a common factor η_{1j}, and the SAT-V variability for coached and uncoahed groups ($\left(u_{3j},\ u_{4j}\right)$) can be explained by a common factor η_{2j}. Formally, we formulate a HLM2 factor (HLM2F) model as follows:

L-1:

$$d_{ij} = \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \varepsilon_{ij}$$

L-2:

$$\beta_{1j} = \gamma_{10} + \gamma_{11}(\textit{Coaching Hours})_j + \eta_{1j} \qquad (5\text{-}18)$$

$$\beta_{2j} = \gamma_{20} + \lambda_{21}\eta_{1j}$$

$$\beta_{3j} = \gamma_{30} + \gamma_{31}(\textit{Coaching Hours})_j + \eta_{2j}$$

$$\beta_{4j} = \gamma_{40} + \lambda_{42}\eta_{2j},$$

where $\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix} \sim N(0, \Psi)$ for $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$. We refer to this model as the reduced 'model a'.

Note that having specified the level-2 errors as Equation (5-16) or (5-17), we structured the τ covariance matrix as

$$D[u_j] = \Lambda D[\eta_j]\Lambda^T = \Lambda\Psi\Lambda^T = \begin{pmatrix} \psi_{11} & \lambda_{21}\psi_{11} & \psi_{12} & \lambda_{42}\psi_{12} \\ \lambda_{21}\psi_{11} & \lambda_{21}^2\psi_{11} & \lambda_{21}\psi_{12} & \lambda_{21}\lambda_{42}\psi_{12} \\ \psi_{21} & \lambda_{42}\psi_{21} & \psi_{22} & \lambda_{42}\psi_{22} \\ \lambda_{42}\psi_{21} & \lambda_{21}\lambda_{42}\psi_{21} & \lambda_{42}\psi_{22} & \lambda_{42}^2\psi_{22} \end{pmatrix}. \qquad (5\text{-}19)$$

Note also that 'model a' in Equation (5-18) is nested within the full model in Equation (5-13). This fact can be shown in a similar way that I did for the 2 by 2 τ matrix case for the infant vocabulary growth data in section 2 of Chapter 5 (see page 62). Another way of finding the constraints on the τ matrix in order for the reduced model (model a) be nested within the full model is obtained by comparing Equation (5-19), the expression of the structured τ created by formulating the HLM2F model (model a), with the unstructured and saturated τ in Equation (5-13), which was formulated by HLM2.

Using this comparison and with a little algebra, we find five constraints on the τ matrix:

$$\tau_{11}\tau_{22} = \tau_{21}^2,\ \tau_{33}\tau_{44} = \tau_{43}^2,$$

$$\tau_{32}\tau_{33} = \tau_{13}\tau_{43},\ \tau_{33}\tau_{41} = \tau_{43}\tau_{31}, \qquad (5\text{-}20)$$

$$\text{and } \tau_{11}\tau_{33}\tau_{42} = \tau_{21}\tau_{43}\tau_{31}.$$

Thus we obtain Equation (5-19) from the saturated τ matrix in Equation (5-13) by defining

$$\psi_{11} = \tau_{11},\ \lambda_{21} = \frac{\tau_{21}}{\tau_{11}},\ \psi_{22} = \tau_{33},\ \lambda_{42} = \frac{\tau_{43}}{\tau_{33}}, \qquad (5\text{-}21)$$

$$\text{and } \psi_{21} = \tau_{31}$$

with these five constraints. Note that the first two constraints in Equation (5-20) correspond to the statements that the squared correlation between u_{1j} and u_{2j} equals to one ($\rho^2_{u_1u_2} = 1$) and the squared correlation between u_{3j} and u_{4j} equals to one ($\rho^2_{u_3u_4} = 1$).

The results of 'model a' is shown in the third column of Table 5.4. To test the model fit, we use a deviance statistic. The deviance test suggests that 'model a' does not fit as well as the full model, $D_1 - D_0 = 615.386 - 559.378 = 56.008$ with 5 (16-11) d.f., P-value < 0.0001, where D_1 is the deviance for model a, and D_0 is the deviance for HLM2 model.

The next step is to try to find a statistically acceptable model. Since the number of parameters is reduced from 16 to 11 and the estimate of τ matrix is almost insufficient rank, the number of parameters of a model that shows a good fit must be between 11 and 16.

One model that satisfies this condition can be obtained by carefully looking at the estimated tau matrix produced by the HLM2 full model. The estimated tau matrix with its correlation matrix form reveals that the upper left 2 by 2 block matrix is closer to singular that the lower right 2 by 2 block matrix because the upper has the correlation 0.972 whereas the lower's one is 0.863. Based on this observation, we consider a model that only u_{1j} and u_{2j} completely share the common variance. Thus we formulate the following model:

L-1:

$$d_{ij} = \beta_{1j} X_{1ij} + \beta_{2j} X_{2ij} + \beta_{3j} X_{3ij} + \beta_{4j} X_{4ij} + \varepsilon_{ij}$$

L-2:

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + \gamma_{11}(\textit{Coaching Hours})_j + \eta_{1j} \\ \beta_{2j} &= \gamma_{20} + \lambda_{21}\eta_{1j} \\ \beta_{3j} &= \gamma_{30} + \gamma_{31}(\textit{Coaching Hours})_j + \eta_{2j} \\ \beta_{4j} &= \gamma_{40} + \eta_{3j}, \end{aligned} \tag{5-22}$$

where $\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix} \sim N(0, \Psi)$ for $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} & \psi_{13} \\ \psi_{21} & \psi_{22} & \psi_{23} \\ \psi_{31} & \psi_{32} & \psi_{33} \end{pmatrix}$. We refer to this model as the reduced 'model b'. Similar argument that we did for 'model a' tells that 'model b' is nested in the HLM2 model, and it is clear that 'model a' is a nested model in 'model b'.

Note that the factor model that we used in 'model b' for the level-2 error vector $u_j = \left(u_{1j},\ u_{2j},\ u_{3j},\ u_{4j}\right)^T$ is

$$u_j = \begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix} = \Lambda \eta_j , \qquad (5\text{-}23)$$

where $\Lambda = \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and $\eta_j = \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix}$.

The results of fitting 'model b' is in column 4 of Table 5.4. The deviance test reveals that 'model b' does not fit as well as the full model, but we marginally reject the null hypothesis at 0.05 level because $D_2 - D_0$ = 568.848 - 559.378 = 9.470 with 3 (16-13) d.f., and the P-value is 0.024, where D_2 is the deviance for model a, and D_0 is the deviance for HLM2 model.

Knowing that 'model b' is rejected though it is marginal, we still continue the investigation of the model that fits to the data. The number of parameters of the model that fits now must be between 13 and 16. One model that satisfies this condition is specified by adding a specificity component to the model, which will be presented in section 1 of Chapter 6. This model adds four specific variance parameters on 'model a', and thus if all of these are statistically significantly different from zero, the number of parameters of the model is 15, which is one smaller that the HLM2 model. To fit the model that has unique specificity is a topic of future study whose model is formulated in section 1 of Chapter 6.

Finally, I would mention that the results and substantial interpretation for the two dimension HLM2 factor model are consistent with those of Kalaian & Raudenbush

(1996), which analyzed the studies that had the pair of SAT-M and SAT-V, where major substantive conclusions were that coaching was more effective for SAT-M than SAT-V, and the between-study unique effect of SAT-M and SAT-V was negatively correlated. However, the current analysis provides a more precise information on what way study variability emerged and it utilized all of the available studies, whereas Kalaian & Raudenbush (1996)'s model required the existence of control group for each study, which will result in 33 studies for this data set (see Table 5.3).

5-4. **Results from the Simulated Data**

The purpose of this analysis is not only to see whether the theory and the computer program that I developed can recover the parameter values well, but also to evaluate the capacity of the methodology to distinguish between true model and alternative incorrect models in the context of large data application. To answer these questions, I generate the artificial data with known parameters.

We consider a situation in which subtests nested within students that was already stated in section 3 of Chapter 3. To summarize the settings, a test consists of 4 subtests, math1, math2, verbal1, and verbal2, and each student, the total of 100 students, takes one form of the test at the first testing occasion and takes a parallel alternative form of the test at the second testing occasion. No missing observations were assumed. Thus we created a situation that can be conceived as subtests are nested within students. Each student has 8 observations for 100 students, with the total observations of 800. We assumed that true scores of subtests for the same domain such as math1 and math2, and verbal1 and verbal2, are perfectly correlated and also assumed that the correlation between mathematical and verbal proficiency was 0.50. Therefore, non-orthogonal two factors, mathematical and verbal proficiency were considered.

The model was formulated as in Equations (3-29) and (3-31), and the parameters in the model were set as $\gamma_{10} = \gamma_{20} = \gamma_{30} = \gamma_{40} = 500$, $\sigma^2 = 25$, $\lambda_1 = 0.8$ $\lambda_2 = 1.2$, $\psi_{11} = 100$, $\psi_{22} = 100$, $\psi_{12} = 50$. It should be noted that the model is an identified model as shown in section 3 of Chapter 3.

Checking the Data that are generated:

We generate the 800 observations in total, 2 observations from each subtest and thus 8 observations for each subject. The descriptive statistics for this data is in the following tables.

Table 5.5 Descriptive Statistics of 4 Generated Outcomes

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
math1	200	500.7896272	10.7851935	472.3574900	536.1797700
math2	200	500.0342420	9.1171437	468.6287800	523.6194500
verb1	200	499.2721306	11.5485774	475.0847300	528.3763000
verb2	200	499.4860767	13.2175895	466.3821800	535.9121100

To see whether the generated data are reasonable with regards to the means and the standard deviations, we let $y_{1ij} \equiv y_{ij|x_{1ij}=1}$, $y_{2ij} \equiv y_{ij|x_{2ij}=1}$, $y_{3ij} \equiv y_{ij|x_{3ij}=1}$, and $y_{4ij} \equiv y_{ij|x_{4ij}=1}$. Then since $y_{1ij} = \beta_{1j} + \varepsilon_{ij}$, $y_{2ij} = \beta_{2j} + \varepsilon_{ij}$, $y_{3ij} = \beta_{3j} + \varepsilon_{ij}$, and $y_{4ij} = \beta_{4j} + \varepsilon_{ij}$, we have $E(y_{1ij}) = \gamma_{10} = 500$, $E(y_{2ij}) = \gamma_{20} = 500$, $E(y_{3ij}) = \gamma_{30} = 500$, and $E(y_{4ij}) = \gamma_{40} = 500$ for the respective means, and

$$s.d.(y_{1ij}) = \sqrt{Var(y_{1ij})} = \sqrt{\psi_{11} + \sigma^2} = \sqrt{125} \cong 11.18,$$

$$s.d.(y_{2ij}) = \sqrt{Var(y_{2ij})} = \sqrt{\lambda_1^2 \psi_{11} + \sigma^2} = \sqrt{89} \cong 9.43,$$

$$s.d.(y_{3ij}) = \sqrt{Var(y_{3ij})} = \sqrt{\psi_{22} + \sigma^2} = \sqrt{125} \cong 11.18, \text{ and}$$

$$s.d.(y_{4ij}) = \sqrt{Var(y_{4ij})} = \sqrt{\lambda_2^2 \psi_{22} + \sigma^2} = \sqrt{169} \cong 13.00 \text{ for each standard deviation.}$$

Comparing these population means and standard deviations with the corresponding sample means and the standard deviations in Table 5.4, we recognize that

the sample statistics of the means and the standard deviations are close to the populations values.

For the second moment property, the sample variance and covariance matrix can be seen in Table 5.6 or as a correlation matrix in Table 5.7.

Table 5.6 Sample Covariance Matrix of 4 Generated Outcomes

Covariance Matrix, DF = 199

	math1	math2	verb1	verb2
math1	116.3203994	74.5273374	47.8142300	59.2826475
math2	74.5273374	83.1223087	29.7039157	41.4890456
verb1	47.8142300	29.7039157	133.3696391	128.1651658
verb2	59.2826475	41.4890456	128.1651658	174.7046711

Table 5.7 Sample Correlation Matrix of 4 Generated Outcomes

Pearson Correlation Coefficients, N = 200
Prob > |r| under HO: Rho=0

	math1	math2	verb1	verb2
math1	1.00000	0.75793 <.0001	0.38388 <.0001	0.41586 <.0001
math2	0.75793 <.0001	1.00000	0.28212 <.0001	0.34429 <.0001
verb1	0.38388 <.0001	0.28212 <.0001	1.00000	0.83963 <.0001
verb2	0.41586 <.0001	0.34429 <.0001	0.83963 <.0001	1.00000

The population covariance matrix can be represented as

$$D\left[\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}\right] = \begin{pmatrix} \psi_{11}+\sigma^2 & \lambda_1\psi_{11} & \psi_{12} & \lambda_2\psi_{12} \\ \lambda_1\psi_{11} & \lambda_1^2\psi_{11}+\sigma^2 & \lambda_1\psi_{12} & \lambda_1\lambda_2\psi_{12} \\ \psi_{12} & \lambda_1\psi_{12} & \psi_{22}+\sigma^2 & \lambda_2\psi_{22} \\ \lambda_2\psi_{12} & \lambda_1\lambda_2\psi_{12} & \lambda_2\psi_{22} & \lambda_2^2\psi_{22}+\sigma^2 \end{pmatrix} = \begin{pmatrix} 125 & 80 & 50 & 60 \\ 80 & 89 & 40 & 48 \\ 50 & 40 & 125 & 120 \\ 60 & 48 & 120 & 169 \end{pmatrix},$$

and as a correlation matrix,

$$\rho\left[\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix}\right] = \begin{pmatrix} 1 & 0.758 & 0.400 & 0.413 \\ 0.758 & 1 & 0.379 & 0.391 \\ 0.400 & 0.379 & 1 & 0.826 \\ 0.413 & 0.391 & 0.826 & 1 \end{pmatrix}.$$

Comparing these expected values with the sample values, the generated data are again reasonable.

Finally we check the distribution of each observed variable, $y_1, y_2, y_3,$ and y_4, using the Shapiro-Wilk test for normality. The results are in Table 5.8.

Table 5.8 Shapiro-Wilk Test for Normality for the Generated Outcomes

Outcome	Shapiro-Wilk Statistic (W)	P-value (Pr. < W)
Math 1	0.9958	0.8432
Math 2	0.9936	0.5480
Verbal 1	0.9881	0.0918
Verbal 2	0.9919	0.3347

All of the P-values computed from the Shapiro-Wilk statistic are greater than .05. That implies that each outcome can be considered to have a normal distribution, which we expect because each outcome was generated from the sum of two independent normal variates.

The above results all indicate that the data were generated accurately as we specified in the model.

Analysis

Now we move to the analysis. Here we formulate a series of identified models discussed in section 3 of Chapter 3, and fit those series of models that only differ in the level-2 variance-covariance structure to the generated data. Thus, all of the following four models, model 0, model 1, model 3, and model 4, have the same level-1 model as written in Equation (3-29). The difference comes in at the level-2 model error structure. The standard HLM2 expresses the level-2 model for this artificial data as,

L-2:

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{2j} \\ \beta_{3j} &= \gamma_{30} + u_{3j} \\ \beta_{4j} &= \gamma_{40} + u_{4j} \end{aligned} \tag{5-24}$$

where

$$\begin{pmatrix} u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \end{pmatrix} \overset{iid}{\sim} N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} & \tau_{13} & \tau_{14} \\ \tau_{12} & \tau_{22} & \tau_{23} & \tau_{24} \\ \tau_{13} & \tau_{23} & \tau_{33} & \tau_{34} \\ \tau_{14} & \tau_{24} & \tau_{34} & \tau_{44} \end{pmatrix} \right).$$

Or in matrix notation,

$$\beta_j = W_j\gamma + u_j,\ u_j \sim N(0, \tau) \tag{5-25}$$

where $\beta_j = \left(\beta_{1j},\ \beta_{2j},\ \beta_{3j},\ \beta_{4j}\right)^T$, $W_j = \mathbf{I}_4$, $\gamma = \left(\gamma_1,\ \gamma_2,\ \gamma_3,\ \gamma_4\right)^T$, and $u_j = \left(u_{1j},\ u_{2j},\ u_{3j},\ u_{4j}\right)^T$.

All of the models deal with factoring u_j and they can be written as in the matrix form

$$u_j = \Lambda \eta_j \tag{5-26}$$

in general. The covariance structure can be written as

$$\tau = \Lambda \Psi \Lambda^T . \tag{5-27}$$

Thus, structuring the τ matrix and adding restrictions on it, all the models can be derived from the standard HLM2 model, which I call here "model 0". That means that all of the following models are nested within model 0.

Model 0:

Model 0 can be specified by a standard HLM model, but if we use the HLM2F formulation, this model can be expressed by setting Λ as 4×4 identity matrix, i.e., $\Lambda = \mathbf{I}_4$ and by setting $\eta_j = \left(\eta_{1j}, \quad \eta_{2j}, \quad \eta_{3j}, \quad \eta_{4j}\right)^T$, and thus by setting Ψ as a 4×4 symmetric variance-covariance matrix.

Model 1:

Model 1 is the model from which we generated the data. Therefore, it is the model that should best fit to the data. Model 1 is obtained by setting $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_1 & 0 \\ 0 & 1 \\ 0 & \lambda_2 \end{pmatrix}$, $\eta_j = \left(\eta_{1j}, \quad \eta_{2j}\right)^T$ in Equation (5-26), and Ψ as a 2×2 variance-covariance matrix in Equation (5-27).

Note that model 1 is nested within model 0, which was already shown at section 3 of Chapter 5 when we analyzed the SAT meta-analysis of coaching effects data (see Equation (5-17).).

Model 2:

Since model 2 was an unidentified as shown in Section 3 of Chapter 3, it will be omitted from the analysis.

Model 3:

Model 3 is specified by setting $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_1 & \lambda_4 \\ 0 & 1 \\ \lambda_3 & \lambda_2 \end{pmatrix}$, $\eta_j = \left(\eta_{1j}, \quad \eta_{2j}\right)^T$ and Ψ as a 2×2 variance-covariance matrix. Note that model 3 is nested within model 0, and then, model 1 is nested within model 3. The fact that model 3 is nested within model 0 can be shown in a similar way that I did for the SAT meta-analysis of coaching effect data in section 3 of Chapter 5 by comparing the structured τ with the unstructured τ. The fact that model 1 is nested within model 3 can be easily obtained by constraining $\lambda_3 = 0$ and $\lambda_4 = 0$.

Model 4:

Model 4 is specified by setting $\Lambda = \begin{pmatrix} 1 \\ \lambda_1 \\ \lambda_3 \\ \lambda_2 \end{pmatrix}$, $\eta_j = \left(\eta_{1j}\right)$ and Ψ as a 1×1 variance-covariance matrix (scalar). Note that model 4 is nested within model 1 as shown in section 3 of Chapter 3 and thus it is the most restrictive model.

To summarize, a series of 4 models that are formulated above has the following nested models structure, Model 4 $\subset$ Model 1 $\subset$ Model 3 $\subset$ Model 0 ('$\subset$' reads 'nested within'). The model number, its characteristics, and the number of parameters estimated in the τ matrix is summarized in Table 5.9.

Table 5.9. Characteristics of the Specified Models[1]

Model	Model Characteristics	# parameters estimated in the Tau matrix
0	Saturated HLM2 model 4 factors	4(5)/2 = 10
3	2 factors cross loading mis-specified model	4 + 3 = 7
1	2 factors simple loading correct model	2 + 3 = 5
4	1 factor mis-specified model	3 + 1 = 4

Results:

The results are shown in Table 5.10. The estimates of fixed effects are very similar across all the models including the standard errors and all the 95% confidence

[1] Note: Model 4 $\subset$ Model 1 $\subset$ Model 3 $\subset$ Model 0.

intervals captures the true values, i.e., $\gamma_{10} = \gamma_{20} = \gamma_{30} = \gamma_{40} = 500$. The large number of iteration of Model 0 indicates that the level-2 variance-covariance is near singular. It was estimated by HLM2 as:

$$\hat{\tau} = \begin{pmatrix} 124.300 & 98.940 & 48.818 & 65.533 \\ 98.940 & 81.295 & 40.195 & 49.139 \\ 48.818 & 40.195 & 94.183 & 112.590 \\ 65.533 & 49.139 & 112.590 & 143.791 \end{pmatrix},$$

and as a correlation matrix,

$$\rho_{\hat{\tau}} = \begin{pmatrix} 1.000 & 0.984 & 0.451 & 0.490 \\ 0.984 & 1.000 & 0.459 & 0.454 \\ 0.451 & 0.459 & 1.000 & 0.967 \\ 0.490 & 0.454 & 0.967 & 1.000 \end{pmatrix},$$

which supports near singularity argument.

The results for Model 1 represent that the estimates capture all of the true parameters in the model within the range of 95% confidence intervals.

Table 5.10. Results of the Analysis for the Simulated Data by Four Models[2]

	Model 0	Model 1	Model 3	Model 4
# iterations until convergence	1309	4	6	5
Fixed effects estimates				
γ_{10}	499.225(1.164)	499.225(1.166)	499.225(1.013)	499.225(1.045)
γ_{20}	499.055(0.960)	499.055(0.963)	499.055(0.855)	499.055(0.893)
γ_{30}	500.200(1.028)	500.2(1.021)	500.2(1.092)	500.2(0.984)
γ_{40}	500.910(1.246)	500.91(1.243)	500.91(1.273)	500.91(1.168)
Factor Loading estimates				
λ_1	N.A.	0.80569(0.0407)	0.86281(0.0472)	0.79521(0.0676)
λ_2	N.A.	1.24486(0.0562)	1.15753(0.0558)	N.A.
λ_3	N.A.	N.A.	0.05810(0.0438)	1.15745(0.0695)
λ_4	N.A.	N.A.	-0.0903(0.0493)	N.A.
λ_5	N.A.	N.A.	N.A.	0.91993(0.0745)
Random effects estimates				
σ^2	23.609(1.669)	25.3519(1.462)	24.0187(1.385)	57.8552(3.091)
τ_{11}	124.300(19.266)	N.A.	N.A.	N.A.
τ_{21}	98.940(14.987)	N.A.	N.A.	N.A.
τ_{22}	81.295(13.193)	N.A.	N.A.	N.A.
τ_{31}	48.818(12.965)	N.A.	N.A.	N.A.
τ_{32}	40.195(10.716)	N.A.	N.A.	N.A.
τ_{33}	94.183(15.012)	N.A.	N.A.	N.A.
τ_{41}	65.532(15.960)	N.A.	N.A.	N.A.
τ_{42}	49.139(13.000)	N.A.	N.A.	N.A.
τ_{43}	112.590(17.079)	N.A.	N.A.	N.A.
τ_{44}	143.792(22.020)	N.A.	N.A.	N.A.
ψ_{11}	N.A.	123.201(14.283)	90.5223(11.071)	80.2888(12.432)
ψ_{12}	N.A.	50.4351(8.190)	46.2939(7.686)	N.A.
ψ_{22}	N.A.	91.5432(10.901)	107.186(12.766)	N.A.
Log-likelihood	-2704.02	-2707.09	-2706.93	-2880.87
# parameters estimated	15	10	12	9
Deviance	5408.032	5414.18	5413.86	5761.74

[2] Note:

1. Cutoff criterion for getting out of the Fisher scoring loop for HLM2F is 'relative change in the squared length of parameter estimates' < 10^{-6}, which is the same value as HLM2, though HLM2 used changes in log-likelihood).
2. ( ) is the standard error
3. The saturated model was tried by HLM2F, but during the iteration, Ψ estimate became negative definite.
4. Model 2 was unidentified and thus was excluded from the analysis.

We focus on the model fit by comparing the deviance statistics. Since all the models are nested, we can perform the deviance test. Since the order of nesting of the models is Model 4 $\subset$ Model 1 $\subset$ Model 3 $\subset$ Model 0, we perform the deviance test as in the table below.

Table 5.11. Several Results for the Tests of the Model Fit

Test	Deviance	d.f.	P-value
Model 3 against Model 0	5.83 (5413.86 - 5408.03)	4 (15-11)	$p = 0.212$
Model 1 against Model 0	6.15 (5414.18 - 5408.03)	6 (15 - 9)	$p = 0.407$
Model 4 against Model 0	353.71 (5761.74 - 5408.03)	7 (15 - 8)	$p < 0.001$
Model 1 against Model 3	0.32 (5414.18 - 5413.86)	2 (11 - 9)	$p = 0.852$
Model 4 against Model 1	347.56 (5761.74 - 5414.18)	1 (9 - 8)	$p < 0.001$

From the table, the deviance test identifies that model 1 is the most appropriate model among the 4 models, which is the model that generated the data.

Finally, in order to evaluate how good the method of estimation shown in Chapter 4 is, 1000 data sets were generated from the model in Equation (3-29) on page 41 and in Equation (3-31) on page 42 and parameter values on page 81. They were analyzed by Model 1, the correct model (see page 43 and page 86).

A 95 % confidence interval on θ, where θ is a generic symbol for any one of the parameters in the model, was constructed for each data set by the form, $\hat{\theta} \pm 1.96 s.e.(\hat{\theta})$, where $s.e.(\hat{\theta})$ is the estimated standard error for $\hat{\theta}$, which was obtained either from the

method of generalized least squares for the fixed effects parameters (see Equation (4-17) on page 54) or from the diagonal element of the information matrix (see Equation (4-15) on page 53). The value of the multiplier of $s.e.(\hat{\theta})$ was chosen as 1.96, which assumes that $\hat{\theta}$ is normally distributed; this assumption can be justified by a reasonable conjecture that since we have a relatively large number of repetitions (1000 times), the normal approximation was appropriate.

Let p be the probability that the confidence interval covers the true parameter value. For each of the 1000 samples a confidence interval was constructed by the above method. We estimate p by the proportion of coverages among the 1000 repetitions, and denote this by $\hat{p}$. For example, for a fixed effect parameter γ_{10}, 945 times out of 1000 repetitions the interval captured the true parameter value (γ_{10} = 500) in the interval. Therefore, the estimated coverage probability of the interval, $\hat{p}$, is 0.945. Similarly, the estimated coverage probability was computed for the other eight parameters in the model and is represented in the second column of Table 5.12.

Table 5.12. Estimated Coverage Probability and Its Confidence Interval for 1000 Simulations

Parameter	Estimated coverage probability ($\hat{p}$)	A 95% confidence interval on p
γ_{10}	0.945	(0.931, 0.959)
γ_{20}	0.950	(0.936, 0.964)
γ_{30}	0.942	(0.927, 0.957)
γ_{40}	0.949	(0.935, 0.963)
λ_{12}	0.938	(0.923, 0.953)
λ_{42}	0.939	(0.924, 0.954)
ψ_{11}	0.945	(0.931, 0.959)
ψ_{21}	0.946	(0.932, 0.960)
ψ_{22}	0.943	(0.929, 0.957)

Based on the point estimate of the coverage probability (p), we can compute an approximate 95% confidence interval on p by $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$, where n is the number of replications. Thus, in our case, n = 1000. Then, for example, an approximate 95% confidence interval on p for γ_{10} is

$(0.945 \pm 1.96 \cdot 0.0072) \cong (0.945 \pm 0.014) = (0.931, 0.959)$.

An approximate 95% confidence interval on p for the other eight parameters in the model was computed in a similar way and is shown in the third column of Table 5.12.

The average estimated coverage probability of the confidence interval is slightly less than 0.95, with each estimated coverage probability being close to this value. All of the 95% confidence intervals on p for every parameter captured the true value, 0.95. These results add credibility to the method developed in this dissertation estimating parameters and standard errors, when the cluster level (level-2) sample size is large enough and the parameter values are not at the boundary of the parameter space.

Chapter 6. Conclusions and Future Directions

6-1. **Conclusions**

In this dissertation, a method that incorporates factor analysis into the level-2 variance-covariance matrix of the hierarchical linear model was discussed. This approach provides several contributions, methodologically and substantively.

6-1-1. Methodological contributions

HLM2F avoids excess simplification of level-2 variance-covariance matrix when some pair(s) of random effects have high correlation. If we specify many regression coefficients as random using the standard HLM, with maximum likelihood (ml) estimated via the EM algorithm, the number of iterations to obtain the estimates gets too large. However, as long as parameter estimates remain in the interior of the parameter space, the increase in the log-likelihood is assured and we can expect convergence in a reasonable amount of time.

Fisher scoring, when combined with EM, can accelerate convergence quite dramatically. When Fisher scoring fails to converge within the parameter space or it fails to show much improvement on each iteration, a possible cause should be that parameter estimates are close to the boundary of the parameter space. Within this boundary estimation problem, there are two possibilities. One is that some variance estimates are close to zero and the other is that some correlations are close to 1 or -1. The former problem can be easily solved by setting the corresponding random level-2 error to zero. One caution needs to be noted, however. Since setting one random level-2 error to zero

implies that all the covariances related to this random error as well as the variance be zero, we need to check whether we can drop the term by executing a multiparameter test such as a deviance test.

Solving the latter problem is one of the main topics of this dissertation. That is, the problem of either failure of convergence or slow convergence, caused by correlations which are close to 1 or -1, was solved by using factor analysis at level-2.

There is another methodological contribution that considering factor analysis in the level-2 variance-covariance matrix provides. That is, this methodology offers a natural framework for modeling. If we don't have a strong theory or prior knowledge to make regression coefficients fixed, we want to make as many of these coefficients as possible be random so that the model is more general such that it allows the regression coefficients to vary among level-2 cluster units. However, if these random coefficients are highly correlated, the HLM2 methodology fails under Fisher scoring or gives a very slow convergence under EM algorithm, even when variance estimates of each random coefficient differ from zero. In this situation, HLM2F is useful because it constrains covariances, whose estimates have come close to the boundary using factor analysis, while allowing variance to differ from zero. Factor analysis achieves this purpose by decomposing the original covariance matrix into a factor loading matrix and a smaller covariance matrix of factors.

Bryk and Raudenbush (1992) provided a guideline of level-1 model building in their Chapter 9 (pp. 201-204). The key issue of the procedure that they discussed was whether we fix the level-1 regression slope or make it random. In this regard, they

suggested using statistical evidence such as point estimates, univariate χ^2 tests, and multivariate deviance tests. Also they mentioned that low estimated reliabilities and the slow rate of convergence for the EM were useful indicators, and diagnostic themselves, for respecifying a random level-1 coefficient as either fixed or nonrandomly varying.

Even after polishing the level-1 model using the above recommended treatment, the slow rate of convergence for the EM may still occur because of highly correlated random coefficients as we have seen in the examples in Chapter 5. Therefore, I propose a slight modification of the level-1 model building procedure. That is,

1) Make as many level-1 regression coefficients random as needed, based on evidence of non-zero variance in the coefficients;
2) If the rate of convergence is very slow, then consider using the HLM2F model. In terms of specifying factor structure, employ substantive theory of the field in question as well as the estimated correlation structure of $\hat{\tau}$, if it is obtained.
3) Further, a more active use of the HLM2F model would be, when a researcher wants to test his/her theory regarding the simultaneous variability of the random effects among the level-2 units, to specify this theory by HLM2F regarding how random effects are related.

This procedure solves a key dilemma that most of the analysts who use multilevel modeling encounter. That is, when analysts don't have a strong theory or prior knowledge to make regression coefficients fixed, they want to make as many of these coefficients random as is possible, but often the data don't have enough information to estimate all of the parameters. The procedure that analysts use to deal with this dilemma is to fix some

of the coefficients that are not their focus, even when they suspect that there is no reason that those coefficients shouldn't vary among level-2 clusters (Bryk & Frank, 1991). HLM2F is a solution of this dilemma.

6-1-2. Substantive contributions

Expanding modeling possibilities

I would argue that the HLM2F methodology will contribute to educational research because this model certainly expands modeling possibilities, while maintaining natural continuity with the standard HLM2. This statement consists of three claims: recovery of $\hat{\tau}$, improvement over current ad hoc procedures, and view of current ad hoc procedures from the HLM2F framework.

Recovery of $\hat{\tau}$

Suppose a researcher wants to study many random effects simultaneously but he/she cannot obtain a stable estimate of τ because of limitations of both the data and the modeling framework. Suppose, however, he/she can run the HLM2F model by reducing the dimensionality of τ. In this situation, the researcher can recover the τ estimate by constructing it from the estimates of Λ and Ψ by $\hat{\tau} = \hat{\Lambda}\hat{\Psi}\hat{\Lambda}^T$. This certainly gives the educational researchers an opportunity to study the large τ-matrix from the limited data he/she has at hand if some of the random effects in fact go together.

Improvement over the current ad hoc procedure

The second claim is that the HLM2F model represents an improvement over the ad hoc procedures currently practiced by educational researchers when they encounter the above situation, i.e., when they want to make many regression coefficients random, but HLM2 does not allow for this. Let's consider the current practice of the researchers who are in this dilemma. To illustrate this claim specifically, recall the Equation (2-1-1), the level-1 model formulated in a hypothetical school effectiveness study.

L-1:

$$Y_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \beta_{2j}x_{2ij} + \beta_{3j}x_{3ij} + \beta_{4j}x_{4ij} + \beta_{5j}x_{5ij} + \beta_{6j}x_{6ij} + \varepsilon_{ij},\ \varepsilon_{ij} \overset{iid}{\sim} N(0,\sigma^2)$$

At level 2, the researcher formulated the model in which all the level-1 coefficients vary randomly among schools after being accounted for by a certain school characteristic W_j because he/she thought that there is no reason to fix some of the coefficients.

L-2:

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}W_j + u_{2j} \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}W_j + u_{3j} \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}W_j + u_{4j} \\
\beta_{5j} &= \gamma_{50} + \gamma_{51}W_j + u_{5j} \\
\beta_{6j} &= \gamma_{60} + \gamma_{61}W_j + u_{6j}
\end{aligned}$$

where $u_j = (u_{0j}, u_{1j}, u_{2j}, u_{3j}, u_{5j}, u_{6j})^T$ is assumed to be distributed as multivariate normal with a mean of $\mathbf{0}$ and a covariance of τ, a 7×7 symmetric matrix. This is a fairly large

model that has 28 unique parameters in the τ matrix. Such a matrix is generally estimable only if the number of level-1 units is quite large.

If the level-1 sample size is not large, the problem that the researcher faces is that the data are insufficient to estimate all the variances and covariances. Following the current practice, the researcher fixes the slopes that are not central to his/her research question, and let the slopes of interest be random; thereafter he/she writes the report. This practice is understandable, but we know that it is not the best way. It is a compromise forced by the data and the limitation of the model.

However, suppose a second researcher with a different theoretical focus analyzes the same data with a different τ specification, producing different results. There could be no way to evaluate the adequacy of the two summaries of evidence.

Using the HLM2F approach, both researchers could estimate a full 7 by 7 τ and produce identical results. Alternative interpretations could then be evaluated in light of a common summary of evidence.

My claim is that even with the same data, we can do a better job by improving the modeling practice, i.e., formulating the HLM2F model with some a priori substantive information or insight that some of the random slopes are highly correlated. Also, when we think about the consequences that the report can produce when used for decision making of educational policy, the impacts that the results induce can be substantial.

View of the current ad hoc procedure from the HLM2F framework

The third claim goes to the fact that the current practice also can be considered as a special case of the HLM2F model. For example, suppose the researcher's interest is on β_{0j} and β_{1j}. Then, the model the researcher would use is

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}W_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11}W_j + u_{1j} \\
\beta_{2j} &= \gamma_{20} + \gamma_{21}W_j \\
\beta_{3j} &= \gamma_{30} + \gamma_{31}W_j \\
\beta_{4j} &= \gamma_{40} + \gamma_{41}W_j \\
\beta_{5j} &= \gamma_{50} + \gamma_{51}W_j \\
\beta_{6j} &= \gamma_{60} + \gamma_{61}W_j.
\end{aligned}$$

This model actually is equivalent to HLM2F model with

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \\ u_{4j} \\ u_{5j} \\ u_{6j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \begin{pmatrix} \eta_{1j} \\ \eta_{2j} \end{pmatrix},$$

which is a case in which the researcher used two latent factors and assigned 1 to (1,1) and (2,2) elements, and 0's to all the rest of the elements of Λ. Thus, the current practice is a part of the HLM2F model framework, where the researcher naively determined the number of factors as two and the values of factor loading matrix as 1's and 0's in the specified position as in the above. If we use the HLM2F model, we can allow the elements of Λ to be unknown, where the current practice fixed these at zero. Thus, the use of HLM2F not only can solve the dilemma that researchers currently face deciding

which slopes they should fix, but it also reminds the researchers of alternative modeling possibilities. Therefore, the HLM2F model not only expands the flexibility of modeling, but also can reflect the researcher's substantive knowledge of the model. This, in turn, means that the HLM2F modeling requires more consideration of substantive theory because the researcher does have to say which element of the level-2 error u_j"go together." Thus, in HLM2F, it is especially important to note that the decision of specifying a factor structure must be made based on interplay between substantive theories and statistical methods.

Substantive interpretation of latent variables

The second contribution of the HLM2F model to educational research from a substantive standpoint involves interpretation of the latent variables which are used at level-2 in HLM2F. HLM2F provides an opportunity to link level-2 random errors which share common latent variables. Those latent variables may be meaningful substantively and can be interpreted, as we have seen in the infant vocabulary growth example, where we found that one and only one latent variable was necessary to describe the differential growth pattern of vocabulary of young children. This result may refine the current existing theory on infant vocabulary growth or motivate the researchers to create a new substantive theory.

Practical interpretability of the results

The third substantive contribution would have to do with interpretation of the results from a practical viewpoint. When we have many intercorrelated random effects, HLM2 results are hard to interpret practically, even in the case when a large Tau matrix is successfully estimated. It is especially difficult to interpret many covariance parameters simultaneously. Therefore we often interpret only the variance components, and may overlook important evidence of shared variance. HLM2F offers an opportunity for analysts to carefully look at the off-diagonal covariance terms to detect the shared variance and to explore latent variables which might have important meanings. If it's possible to make a concise summary by reducing the dimensionality and by using latent variables, a picture of what kind of factors influence the variability among the level-2 units then emerges. HLM2F provides this possibility.

Overall, the proposed HLM2F model creates new flexibility by integrating two major statistical methodologies, HLM and confirmatory factor analysis. In terms of how to use theory and evidence to formulate models, more work needs to be undertaken.

6-1-3. Unsolved Methodological Problems

Confirmation of global maximum

When maximum likelihood is chosen as a method of estimation, there is always a possibility that the likelihood function has multiple maxima, or a flat likelihood in the neighborhood of the maximum likelihood estimate. Thus, it is a good idea to graph the likelihood or the log-likelihood to detect this problem. However, this is difficult since we

have many parameters and we want to sketch the log-likelihood curve, which is a function of many parameters, while the standard graph is limited to three-dimensions. To address this issue, one might consider using the profile likelihood. This method is appealing because we can visually see the shape of the likelihood as the function of the target parameter by taking into account other estimates of parameters. The profile likelihood can be defined (see for example, Gorthwaite, Jolliffe, & Jones, 1995) as follows:

Definition: *Profile likelihood*

Consider a vector of parameters $\theta = (\theta_1, \theta_2)$, with likelihood function $L(\theta_1, \theta_2; \mathbf{y})$, where $\mathbf{y}$ is an observed vector which comes from a density $f_{\mathbf{Y}}(\mathbf{y}; \theta)$. Suppose that $\hat{\theta}_{2\cdot 1}$ is the MLE of θ_2 for a given value of θ_1, then the profile likelihood for θ_1 is $L(\theta_1, \hat{\theta}_{2\cdot 1}; \mathbf{y})$.

The method of computing the profile likelihood involves the following steps:

1) Fix a parameter θ_1, in which you are interested, to some specific value, e.g., $\theta_1^{(0)}$.

2) Obtain the MLE of θ_2, conditional on the fixed value of the target parameter, θ_1.

3) Compute the likelihood $L(\theta_1, \theta_2; \mathbf{y})$ by plugging those values in the likelihood formula, i.e., obtain $L(\theta_1 = \theta_1^{(0)}, \theta_2 = \hat{\theta}_{2\cdot 1}; \mathbf{y})$.

4) Change the value of the target parameter θ_1 and repeat the steps 1) ~ 3) for all possible θ_1.

5) Draw the profile likelihood curve against θ_1.

One technical difficulty arises when we consider the implementation of this methodology. That is, in step 2, we need to compute the MLE for a given value of θ_1. If θ_2 involves a variance component and we choose a bad value for θ_1, Fisher scoring fails by providing a negative estimate. In our HLM2F model in Equation (3-9) (or Equation (A-1) in Appendix), $\theta = (\gamma, \phi)$,where $\phi = (\lambda, \psi, \sigma^2)$ (see Equation (A-9)). Suppose, for example, $\lambda = (\lambda_1, \lambda_2)^T$ and $\psi = (\psi_{11}, \psi_{21}, \psi_{22})^T$. Suppose also that we are interested in the behavior of ψ_{11}. Then, in this case, $\theta_1 = \psi_{11}$, and $\theta_2 = (\gamma^T, \lambda^T, \psi_{21}, \psi_{22})$. To compute the profile likelihood of ψ_{11}, we compute the MLE of θ_2 given ψ_{11}. But if we fix ψ_{11} at an unlikely value, then Fisher scoring produces an unacceptable MLE of θ_2, e.g., $\hat{\psi}_{22}$ becomes a negative value, etc. Thus, as long as we use Fisher scoring, we will face this problem in computing the profile likelihood.

A solution for this might be to use the EM algorithm. The EM algorithm assures convergence to a local maximum and thus we can compute the profile likelihood at any point evaluated, at least, at the local MLE. This certainly motivates one to develop an EM algorithm for the HLM2F model, but whether it is the global maximum or not is still in question.

A non-graphical way of checking global maximum was developed by Gan and Jiang (1999). Their claim is that a global maximizer would satisfy

$$\left(\frac{\partial l}{\partial \theta}\right)^2 + \frac{\partial^2 l}{\partial \theta^2} \approx 0,$$

which is consistent with the property that at the global MLE, an equality to derive the Cramer-Rao lower bound should be satisfied,

$$E_{\theta_0}\left(\frac{\partial l}{\partial \theta}|_{\theta_0}\right)^2 + E_{\theta_0}\left(\frac{\partial^2 l}{\partial \theta^2}|_{\theta_0}\right) = 0,$$

where l is the log-likelihood function, θ is an unknown parameter and θ_0 is the true θ. Based on the observation on the above approximate equality, they developed a large-sample test that is practically usable. Their Monte Carlo studies on distributions that involve local maxima, including a normal mixture distribution, showed that the observed significance level was close to α level and the power was very high in a sample size of 500 in a simple random sampling.

Identification

Necessary and sufficient conditions for model identification are still an unsolved question when we use the factor analytic model. There are no sufficient conditions known. In practice, what we can do now is, once a model satisfies the necessary conditions, try to run the software and see if the software can produce the estimates. Then we might try different starting values for the estimates of the parameters and see if the software can produce the same estimates. If it does, we can have a little confidence that the model is identified, though it is not mathematically proved. This practice is based on the concept of local identification, which means that the information matrix is invertible at the MLE.

Wald (1950) provided an alternative sufficient condition for local identification using a Jacobian determinant which is known as Wald's rank rule (Bollen, 1989). The covariance equations that need to be solved are non-linear simultaneous equations in

terms of parameters. A general way of solving non-linear equations is an iterative algorithm using the Jacobian of the simultaneous equations. Wald's suggestion is that if the Jacobian has non-zero determinant at the value we plugged in to the equations, then the parameter is identified at that value. Since these days symbolic computation of a determinant is possible using a software package such as Mathematica (Wolfran (1991)), Wald's suggestion could be extended to a statement of necessary and sufficient condition of model identification, such that if the Jacobian determinant is not zero at the arbitrary point in the parameter space, then the model is identified, and vice versa.

6-2. **Future Directions**

The issues of confirmation of global maximum and model identification, which were mentioned in the section of unsolved problems, are more general statistical problems that apply to many other models than HLM2F which I developed in this dissertation. Clearly more studies on these topics are needed to solve these problems.

In this section, I mainly discuss extensions of the HLM2F model. Since correlated variables are common in social and behavioral sciences, a variety of the extensions of HLM2F can be considered. Those include:

a. Including a specificity parameter;

b. Multivariate hierarchical linear model with factor structure (MHLMF), where factor analysis is applied to the multivariate HLM model;

c. Nonlinear hierarchical generalized linear model with factor structure (HGLMF), where factor analysis is applied to the two-level hierarchical generalized linear model

(HGLM). When the outcome variables are dichotomous responses to test items, it can be shown that HGLMF is equivalent to a multi-dimensional two-parameter (2-P) IRT model. This is a confirmatory item factor analysis.

6-2-1. Including specificity parameters

Including specificity parameters into HLM2F is straight forward and is useful. As I suggested in section 5.3, this model is immediately applicable to the SAT meta-analysis data.

Model:

L-1:

$$Y_j = X_j\beta_j + r_j, \tag{6-2-1.1}$$

where Y_j is a $n_j \times 1$ vector, X_j is a $n_j \times R$ matrix, β_j is a $R \times 1$ vector, and r_j is a $n_j \times 1$ vector, and

$$r_j \sim N(0, \sigma^2 \mathbf{I}_{n_j}) \tag{6-2-1.2}$$

where $\mathbf{I}_{n_j}$ denotes an identity matrix of size n_j.

L-2:

$$\beta_j = W_j\gamma + u_j \tag{6-2-1.3}$$

$$u_j = \Lambda\eta_j + \xi_j \tag{6-2-1.4}$$

where W_j is a $R \times F$ matrix, γ is a $F \times 1$ vector, u_j is a $R \times 1$ vector, Λ is a $R \times M$ matrix, η_j is a $M \times 1$ vector $(R \geq M)$, and ξ_j is a $R \times 1$ diagonal vector, and

$$u_j \sim N(0, \tau), \tag{6-2-1.5}$$

$$\eta_j \sim N(0, \Psi), \tag{6-2-1.6}$$

$$\xi_j \sim N(0, \Omega), \tag{6-2-1.7}$$

where

$$\Omega = diag(\omega_1, \omega_2, \ldots, \omega_R). \tag{6-2-1.8}$$

Thus, we have

$$\tau = \Lambda \Psi \Lambda^T + \Omega. \tag{6-2-1.9}$$

and the total variance $D[Y_{ij}] \equiv \Sigma$, a $R \times R$ matrix, is decomposed into

$$\begin{aligned} \Sigma &= \tau + \sigma^2 \mathbf{I}_{n_j} \\ &= \Lambda \Psi \Lambda^T + \Omega + \sigma^2 \mathbf{I}_{n_j}. \end{aligned} \tag{6-2-1.10}$$

In factor analysis terms, the first term in Equation (6-2-1.10) is called the communality (common factor variance), and the second term is called the specificity (part of a test's true variance which is not shared with any other tests in a battery) (Thurston, 1947). Note that since we are formulating a factor model at level-2, there is no error variance left at level-2 variance-covariance τ. That is, the true variance τ is decomposed to communality and uniqueness by the level-2 model. In his factor analysis model, Thurston (1947, Chapter 3) actually considered the model in terms of variance decomposition as follows:

$$\text{Total Variance} = \text{True Score Variance} + \text{Error Score Variance}$$

$$= \underbrace{\text{Communality} + \text{Specificity}}_{\text{True Score Variance}} + \text{Error Variance} \qquad (6\text{-}2\text{-}1.11)$$

$$= \text{Communality} + \underbrace{\text{Specificity} + \text{Error Varinace.}}_{\text{Uniqueness}}$$

Usually in factor analysis, the equation that is represented by the third decomposition in (6-2-1.11) is used. But by using the model represented by Equation (6-2-1.1) to (6-2-1.10), we are able to further decompose the uniqueness into the specificity and the error variance.

Note that if we constrain $\omega_r = 0$ for all r, then it reduces to the proposed model in Chapter 3.

In section 6-2-2, I will present the model without the specificity term, i.e., $\xi_j = 0$ for all j, because in a technical sense, the model seems more useful in solving the nonconvergence problem when some of the correlation estimates are close to 1 or -1, although inclusion of the specificity term is straightforward. Then, in section 6-2-3, I will show that the 2-Parameter Item Response Theory (IRT) model can be formulated by a nonlinear version of the HLM2F model.

6-2-2. Multivariate HLM with a Factor Model (2-level Multivariate Model)

General Formulation of the Model

Suppose the complete data consists of P observations. For example, if a test consists of Pth subtests and the examinee responds to all the subtests, then we have P

outcomes for that subject. In a longitudinal study, if *P*th waves of observations are planned, and if we have succeeded to follow up, then we have complete *P* waves of observations. Often in these settings, we have missing observations. However, if we can assume that missing observations occurred at random (MAR), we can use all the information we have gotten to estimate the complete data model without bias. To develop a general multivariate model, we utilize the idea of Jennrich and Schluchter (1986) and Thum (1997). That is, we formulate a level-1 model by two steps. The first step of the level-1 model links observed outcome and complete outcome, which is

L-1:

$$Y_{ij} = \sum_{p=1}^{P} M_{ipj} Y_{pj}^{*} \qquad (6\text{-}2\text{-}2.1)$$

where Y_{ij} is a scalar and *i*th observation of *j*th level-2 unit, Y_{pj}^{*} is a scalar and p th observation of level-1 unit, and M_{ipj} is an indicator and is 1 if *i*th observation is the observation of p th level-1 unit, 0 if otherwise. Thus the matrix M_{ipj} indicates which observation of complete data we have got, or it tells us the missing pattern of the observations. In matrix notation, we can write

$$Y_j = M_j Y_j^{*}, \qquad (6\text{-}2\text{-}2.2)$$

where Y_j is a $p_j \times 1$ vector of observed data outcome, M_j is a $p_j \times P$ matrix of missing pattern, and Y_j^{*} is a $P \times 1$ vector of complete data observation of the outcome. Then, we formulate the second part of the level-1 model, which is the standard formulation of the

level-1 model in HLM. Thus, we formulate the standard model for the complete data, not for the observed data.

$$Y_j^* = X_j^* \beta_j + r_j^*, \; r_j^* \sim N(0, \Sigma), \tag{6-2-2.3}$$

where Σ is the $P \times P$ variance-covariance matrix for r_j^*, i.e., $\Sigma = D[r_j^*] = D[Y_j^* | \beta_j]$. Plugging Equation (6-2-2.3) into Equation (6-2-2.2), we obtain the level-1 model by letting $X_j = M_j X_j^*$ and $r_j = M_j r_j^*$.

L-1:

$$Y_j = X_j \beta_j + r_j, \; r_j \sim N(0, \Sigma_j) \tag{6-2-2.4}$$

where

$\Sigma_j = M_j \Sigma M_j^T$, a $p_j \times p_j$ symmetric matrix. Notice that now we have the subscript j in Σ that was induced by a missing indicator matrix M_j and that Σ_j is actually a submatrix of Σ, a $P \times P$ symmetric matrix. Thus, what multiplying M_j and M_j^T from the left and the right does is to pick up the subset of the Σ matrix. The level-2 model is a standard formulation.

L-2:

$$\beta_j = W_j \gamma + u_j, \tag{6-2-2.5}$$

Now we consider the case when $u_j = \Lambda \eta_j$, then we have the level-2 model

$$\beta_j = W_j \gamma + \Lambda \eta_j, \eta_j \sim N(0, \Psi). \tag{6-2-2.6}$$

If we put Equation (6-2-2.2), (6-2-2.3), and (6-2-2.6) together, we obtain the combined model.

$$Y_j = M_j(X_j^* W_j \gamma + X_j^* \Lambda \eta_j + r_j). \qquad (6\text{-}2\text{-}2.7)$$

We write this general two-level multivariate linear model, which has a factor structure at level-2, with a slightly more general form:

$$\begin{aligned} Y_j &= M_j(A_{1j}\theta_1 + A_{2j}\Lambda\eta_j + r_j) \\ &= M_j A_{1j}\theta_1 + M_j A_{2j}\Lambda\eta_j + M_j r_j, \end{aligned} \qquad (6\text{-}2\text{-}2.8)$$

where $r_j \sim N(0, \Sigma)$, $\eta_j \sim N(0, \Psi)$. If we let $X_j = M_j A_{1j}$, $A_j = M_j A_{2j}$, and $e_j = M_j r_j$, then we have a standard form of HLM with level-2 factor structure as in Equation (3-9),

$$Y_j = X_j\theta_1 + A_j\Lambda\eta_j + e_j, \qquad (6\text{-}2\text{-}2.9)$$

where $e_j \sim N(0, \Sigma_j)$ for Σ_j is a $p_j \times p_j$ symmetric matrix and $\Sigma_j = M_j \Sigma M_j^T$. Heterogeneity of variance (Note the suffix for Σ) appears because of varying missing pattern matrices M_j, and Σ_j is a subset of Σ.

Now we show two examples to illustrate how to apply the above general formulation to the specific cases.

Example 1. Growth Model

Suppose that the same math test was administered for *n* elementary school students from 1st grade to 5th grade, but some of the students did not take the test at some occasions of testing. We wish, for example, to know how math proficiency of the students grows over time and what the variation of the proficiency among the students is. Then, in a general formulation of the multivariate model, Y_j^*, complete data for student *j*,

is $Y_j^* = (Y_{1j}^*, Y_{2j}^*, Y_{3j}^*, Y_{4j}^*, Y_{5j}^*)^T$, a 5×1 vector in Equation (6-2-2.2), and the number of occasions $P = 5$. To be specific, if student 1 took the test at all of the grades, then we have

$$\begin{pmatrix} Y_{11} \\ Y_{21} \\ Y_{31} \\ Y_{41} \\ Y_{51} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_{11}^* \\ Y_{21}^* \\ Y_{31}^* \\ Y_{41}^* \\ Y_{51}^* \end{pmatrix}. \qquad \text{(6-2-2.10)}$$

If student 2 missed the third occasion, then we have

$$\begin{pmatrix} Y_{12} \\ Y_{22} \\ Y_{32} \\ Y_{42} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} Y_{11}^* \\ Y_{21}^* \\ Y_{31}^* \\ Y_{41}^* \\ Y_{51}^* \end{pmatrix}. \qquad \text{(6-2-2.11)}$$

To formulate the complete data level-1 model in Equation (6-2-2.3), suppose we decide to use a quadratic function for modeling mean growth for student j, and the grade was coded as (1st grade, 2nd grade, 3rd grade, 4th grade, 5th grade) = (-2, -1, 0, 1, 2), which centers the age variable around 3rd grade. Then, the parameter β_j is a 3×1 vector $\beta_j = (\beta_{0j}, \beta_{1j}, \beta_{2j})^T$ and the design matrix X_j^* is

$$X_j^* = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix}$$ for all students. Thus, the whole level-1 model for complete data is

$$\begin{pmatrix} Y_{11}^{*} \\ Y_{21}^{*} \\ Y_{31}^{*} \\ Y_{41}^{*} \\ Y_{51}^{*} \end{pmatrix} = \begin{pmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{pmatrix} \begin{pmatrix} \beta_{0j} \\ \beta_{j} \\ \beta_{2j} \end{pmatrix} + \begin{pmatrix} r_{1j}^{*} \\ r_{2j}^{*} \\ r_{3j}^{*} \\ r_{4j}^{*} \\ r_{5j}^{*} \end{pmatrix}, \begin{pmatrix} r_{1j}^{*} \\ r_{2j}^{*} \\ r_{3j}^{*} \\ r_{4j}^{*} \\ r_{5j}^{*} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right) \qquad (6\text{-}2\text{-}2.12)$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 \end{pmatrix}$.

This is an unconstrained (or saturated) level-1 covariance. Modeling this covariance structure using a smaller number of parameters by giving restrictions such as homogeneous case, heterogeneous but independent case, a first order autoregressive model (AR(1)) case, factor analysis (exploratory model) case, and so forth, were mentioned in Jennrich and Schluchter (1986) and Thum (1997).

For the level-2 model in Equation (6-2-2.5), suppose that we found a high negative correlation between residual intercept u_{0j} and the residual slope u_{1j}, but no correlation between u_{0j} and the residual rate of acceleration u_{2j}, and between u_{1j} and u_{2j}, even after we included all of the necessary predictors that could explain individual differences such as gender, race, socio-economic status (SES), then we might consider the model in Equation (6-2-2.6). That is, it might be useful to think about

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix}, \begin{pmatrix} \eta_{0j} \\ \eta_{1j} \end{pmatrix} \sim N(\mathbf{0}, \Psi) \qquad (6\text{-}2\text{-}2.13)$$

where $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$.

This factor model structures the τ as

$$\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} \\ \tau_{10} & \tau_{11} & \tau_{12} \\ \tau_{20} & \tau_{21} & \tau_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ \lambda & 0 \\ 0 & 1 \end{pmatrix}^T = \begin{pmatrix} \psi_{11} & \lambda\psi_{11} & 0 \\ \lambda\psi_{11} & \lambda^2\psi_{11} & 0 \\ 0 & 0 & \psi_{22} \end{pmatrix} \qquad (6\text{-}2\text{-}2.14)$$

and reduces the number of parameters in τ from 6 to 3.

Example 2. Multivariate Outcome Growth Model

Suppose that students in primary school took 2 tests such as Reading and Mathematics each year from 1st grade to 3rd grade. In this type of study, it is natural that we have missing observations for either or both tests at some time points for some students. We assume that these missing patterns are missing at random (MAR). In this setting, the complete data for each student is $Y_j^* = (Y_{1j}^*, Y_{2j}^*, Y_{3j}^*, Y_{4j}^*, Y_{5j}^*, Y_{6j}^*)^T$ where Y_{1j}^* is the reading score for student j at grade 1, Y_{2j}^* is the mathematics score for student j at grade 1, Y_{3j}^* is the reading score for student j at grade 2, Y_{4j}^* is the mathematics score for student j at grade 2, Y_{5j}^* is the reading score for student j at grade 3, Y_{6j}^* is the mathematics score for student j at grade 3. To formulate the complete data level-1 model in Equation (6-2-2.3), suppose we decide to use linear function for the mean growth for student j and the grade was coded as (1st grade, 2nd grade, 3rd grade) = (-1, 0, 1), which centers the age variable around 2nd grade. Then, the parameter β_j is a 4×1 vector $\beta_j = (\beta_{0j}, \beta_{1j}, \beta_{2j}, \beta_{3j})^T$, where β_{0j} is the mean reading intercept for student j at grade 2,

β_{1j} is the mean reading slope for student j, β_{2j} is the mean mathematics intercept for student j at grade 2, β_{3j} is the mean mathematics slope for student j. The design matrix X_j^* in Equation (6.2.2.3) is

$$X_j^* = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$ for all students. Thus, the whole level-1 model for complete data is

$$\begin{pmatrix} Y_{1j}^* \\ Y_{2j}^* \\ Y_{3j}^* \\ Y_{4j}^* \\ Y_{5j}^* \\ Y_{6j}^* \end{pmatrix} = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \beta_{0j} \\ \beta_{1j} \\ \beta_{2j} \\ \beta_{3j} \end{pmatrix} + \begin{pmatrix} r_{1j}^* \\ r_{2j}^* \\ r_{3j}^* \\ r_{4j}^* \\ r_{5j}^* \\ r_{6j}^* \end{pmatrix}, \begin{pmatrix} r_{1j}^* \\ r_{2j}^* \\ r_{3j}^* \\ r_{4j}^* \\ r_{5j}^* \\ r_{6j}^* \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \Sigma \right) \qquad (6\text{-}2\text{-}2.15)$$

where $\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} & \sigma_{15} & \sigma_{16} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} & \sigma_{25} & \sigma_{26} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} & \sigma_{35} & \sigma_{36} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 & \sigma_{45} & \sigma_{46} \\ \sigma_{51} & \sigma_{52} & \sigma_{53} & \sigma_{54} & \sigma_5^2 & \sigma_{56} \\ \sigma_{61} & \sigma_{62} & \sigma_{63} & \sigma_{64} & \sigma_{65} & \sigma_6^2 \end{pmatrix}$.

Thus Σ matrix, i.e., within-student variance-covariance, is 6×6 in this case. In addition to the ways of structuring the covariance matrix mentioned in example 1., we can think about other patterns here because there is a content area factor, i.e., reading vs

mathematics, as well as a time factor within students. For example, $\Sigma = \begin{pmatrix} \Delta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Delta \end{pmatrix}$,

where $\Delta = \begin{pmatrix} \delta_1^2 & \delta_{12} \\ \delta_{21} & \delta_2^2 \end{pmatrix}$ and $\mathbf{0}$ is a 2×2 null matrix, represents a block homogeneous

variance-covariance pattern which corresponds to $\Sigma = \sigma^2 \mathbf{I}$, a homogeneous independent variance-covariance for example 1. The within-subject variance-covariance

pattern $\Sigma = \begin{pmatrix} \Delta & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \Delta & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \Delta \end{pmatrix}$ implies that reading and mathematics scores are correlated at each

time point within students as in Δ, and the within-student variance covariance is constant over three time points, but has no correlation with other time points.

Now for the level-2 model, suppose that we found a high positive correlation between the residual reading intercept u_{0j} and the residual mathematics intercept u_{2j}, and between the residual reading slope u_{1j} and the residual mathematics slope u_{3j}, but no correlation between u_{0j} and u_{1j}, between u_{0j} and u_{3j}, between u_{1j} and u_{2j}, and between u_{2j} and u_{3j} even after being accounted for by all of the necessary predictors such as gender, race, socio-economic status (SES) for example, that could explain individual differences. Then we might consider the model in Equation (6-2-2.6). That is, it might be useful to think about

$$\begin{pmatrix} u_{0j} \\ u_{1j} \\ u_{2j} \\ u_{3j} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \eta_{0j} \\ \eta_{hj} \end{pmatrix}, \begin{pmatrix} \eta_{0j} \\ \eta_{hj} \end{pmatrix} \sim N(\mathbf{0}, \Psi) \qquad (6\text{-}2\text{-}2.16)$$

where $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$.

This factor model structures the τ as

$$\begin{pmatrix} \tau_{00} & \tau_{01} & \tau_{02} & \tau_{03} \\ \tau_{10} & \tau_{11} & \tau_{12} & \tau_{13} \\ \tau_{20} & \tau_{15} & \tau_{22} & \tau_{23} \\ \tau_{30} & \tau_{31} & \tau_{32} & \tau_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}^T$$

$$= \begin{pmatrix} \psi_{11} & \psi_{12} & \lambda_1\psi_{11} & \lambda_2\psi_{12} \\ \psi_{21} & \psi_{22} & \lambda_1\psi_{21} & \lambda_2\psi_{22} \\ \lambda_1\psi_{11} & \lambda_1\psi_{12} & \lambda_1^2\psi_{11} & \lambda_1\lambda_2\psi_{12} \\ \lambda_2\psi_{21} & \lambda_2\psi_{22} & \lambda_1\lambda_2\psi_{21} & \lambda_2^2\psi_{22} \end{pmatrix} \qquad (6\text{-}2\text{-}2.17)$$

and reduces the number of parameters in τ from 10 to 5.

6-2-3. Application to Item Response Theory Model

The purpose of presenting an alternative formulation for a unidimensional 2-parameter Item Response Theory (IRT) model here is that, since the reformulation makes the IRT model a special case of nonlinear mixed model with factor structure, it gives us an opportunity to examine the model and the theory, from a standard statistical point of view. Also, the reformulation gives us a model that is easier to extend to a multidimensional IRT model and to multiple levels of nesting.

We will first give the perspective of IRT as a nonlinear factor analysis. What we mean by "nonlinear" here is a nonlinear transformation of the expected value of the outcome is modeled by a linear combination of parameters and independent variables. From this view, we will show that if we adopt the IRT as a nonlinear factor analysis perspective, it is straightforward to extend the usual unidimensional IRT model to a multidimensional IRT model. Though the current IRT model, whether uni-dimensional or multi-dimensional, is used in the exploratory factor analysis mode, it will be shown that a confirmatory mode non-linear factor analysis can be executed without much difficulty, if we formulate the IRT model as a hierarchical generalized linear model (HGLM) (See Chapter 5 & 6, Bryk, Raudenbush, & Congdon (1996)).

Usually a test consists of many items, and the items are supposed to measure the student's proficiency. Then we want to know the current status of the proficiency of the examinee and how his/her proficiency grew over time if the longitudinal data are available. The key for the modeling here is to represent the IRT model as the HGLM level-1 model for a nonlinear factor analysis, and if we use dummy variables, the same

model can be represented by HGLM with level-2 factor structure. Suppose that a test that consists of n dichotomous items is administered for J students, whose response Y_{ij} is scored 1 if correct, 0 if not, and suppose that J students are a random sample from the student population. The distribution of random variable Y_{ij} conditional on the probability of correct response of student j for item i (denoted as p_{ij}) is Bernoulli with mean p_{ij} ($0 \le p_{ij} \le 1$) i.e., $\Pr(Y_{ij} = 1 | p_{ij}) = p_{ij}$. If we write this data generation process in a regression form, then the sampling part of the level-1 model (the unit is item) is:

L-1 Sampling Model:

$$Y_{ij} = p_{ij} + \varepsilon_{ij}, \tag{6-2-3.1}$$

where $\varepsilon_{ij} \sim Ber(0, p_{ij}(1-p_{ij}))$ and the ε_{ij}'s are independent from each other. Note that the conditional mean and the conditional variance of the outcome variable Y_{ij} are

$$E(Y_{ij} | p_{ij}) = p_{ij}, \text{ and } Var(Y_{ij} | p_{ij}) = p_{ij}(1-p_{ij}). \tag{6-2-3.2}$$

Now, for the model for the mean p_{ij}, since the range of p_{ij} is restricted between 0 and 1, and in order for the transformed variable to have the range theoretically -∞ to +∞, we make a logit transformation, i.e.,

$$\eta_{ij} = \log(\frac{p_{ij}}{1-p_{ij}}) \equiv \log it(p_{ij}), \tag{6-2-3.3}$$

and then we assume that η_{ij} has a normal distribution. For notational convenience, we write the logit inverse transformation as

$$p_{ij} \equiv \text{logit}^{-1}(\eta_{ij}), \text{ where } p_{ij} = \frac{1}{1+\exp(-\eta_{ij})}. \tag{6-2-3.4}$$

Now, the level-1 structural model for the 2-parameter IRT model for item i of student j would be:

L-1 Structural Model:

$$\eta_{ij} = \lambda_i(\theta_j - \delta_i). \tag{6-2-3.5}$$

Combining the level-1 sampling and structural models, we obtain the Level-1 model as:

L-1 Model:

$$Y_{ij} = \text{logit}^{-1}\left[\lambda_i(\theta_j - \delta_i)\right] + \varepsilon_{ij}, \tag{6-2-3.6}$$

where

$$\varepsilon_{ij} \sim Ber(0, p_{ij}(1 - p_{ij})) \text{ for } 0 \le p_{ij} \le 1. \tag{6-2-3.7}$$

and λ_i is the item discrimination parameter for item i, θ_j is person j's ability (proficiency), δ_i is the ith item's difficulty parameter. The scaling constant $D \cong 1.7$ that makes the logistic model comparable to the normal threshold model is omitted from the model representation without loss of generality. Note that if we consider all the λ_i's to be the same for all the items, then it will be a Rasch model.

ANOVA type formulation of 2-P IRT using HGLM

We now show that the inside of the inverse logit function of Equation (6-2-3.6) can be written in the same format as the linear factor analysis model. Specifically, let's consider that we have only three items in the test. Then the level-1 structural model of the 2-parameter IRT model can be written as

L-1 Structural Model:

$$\begin{aligned} &\textit{Item 1: } \eta_{1j} = \lambda_1(\theta_j - \delta_1) \\ &\textit{Item 2: } \eta_{2j} = \lambda_2(\theta_j - \delta_2), \\ &\textit{Item 3: } \eta_{3j} = \lambda_3(\theta_j - \delta_3) \end{aligned} \tag{6-2-3.8}$$

where $\theta_j \sim N(0,1)$. The variance of θ_j was fixed to 1 in order to identify all of the item discrimination parameters. In matrix format, Equation (6-2-3.8) can be written as

$$\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix} = \begin{pmatrix} -\lambda_1\delta_1 \\ -\lambda_2\delta_2 \\ -\lambda_3\delta_3 \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}\left(\theta_j\right) = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix}\left(\theta_j\right), \tag{6-2-3.9}$$

where

$$\mu_i \equiv -\lambda_i\delta_i \text{ for all } i. \tag{6-2-3.10}$$

Or, we can write it in matrix form as

$$\eta_j = \mu + \Lambda\theta_j. \tag{6-2-3.11}$$

Then, at level-2 whose unit is student, we have,

$$\theta_j = \beta_{00} + u_j, \tag{6-2-3.12}$$

where $u_j \sim N(0,1)$. Note that the $Var(u_j) = 1$ corresponds to $Var(\theta_j) = 1$.

Thus the logit-transformed mean vector η_j conditional on θ_j has the same structure as the conditional mean of the standard one-population factor analysis model, i.e., $E(Y|\theta) = \mu + \Lambda\theta$. Remember that the key assumption of the linear factor analysis is that given the factor score (latent ability), the Y's (outcome variables) are independent. In the IRT case, the local independence assumption (that, given θ_j, the observations Y_{ij}'s are independent) corresponds to the conditional independence assumption for the linear

factor analysis model. The difference between the linear factor analysis model and IRT model is whether the link function is identity or logit. Thus, we can view the IRT model as a non-linear factor analysis model. If we do so, then the item difficulty parameter is the mean of the logit, and the item discrimination parameters are contained in the factor loadings matrix Λ.

Regression type formulation of 2-P IRT using HGLM

Rasch (1-P IRT) model

Though the above formulation is useful for recognizing that the IRT model involves a factor structure, it does not facilitate finding the estimation procedure similar to the one which is executed by the current HGLM. In order to do that, we reformulate the above IRT model by a hierarchical model with dummy variables, which was introduced by Kamata (1998) and Cheong & Raudenbush (1999), who applied this model to the 1-parameter (1-P) IRT (Rasch) model. In the following presentation, since the level-1 sampling model is always the same, i.e., Bernoulli distribution, I will omit it.

To facilitate the idea of representing a 2-parameter (2-P) IRT model by HGLM, we use no-intercept model first. As before, we have n items and J students. For item i and student j,

L-1 Structural Model:

$$\begin{aligned}\eta_{ij} &= \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \beta_{3j}X_{3ij} + \cdots + \beta_{qj}X_{qij} + \cdots + \beta_{n}X_{nij} \\ &= \sum_{q=1}^{n}\beta_{qj}X_{qij}\end{aligned} \qquad (6\text{-}2\text{-}3.13)$$

where X_{qij} is the indicator variable and is 1 if $q = i$ and 0 if $q \neq i$ for all i. At level-2, all the coefficients randomly vary among students, but with the same u_{1j}.

L-2:

$$\begin{aligned} \beta_{1j} &= \gamma_{10} + u_{1j} \\ \beta_{2j} &= \gamma_{20} + u_{1j} \\ &\cdots \\ \beta_{qj} &= \gamma_{q0} + u_{1j} \\ &\cdots \\ \beta_{nj} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{6-2-3.14}$$

where $u_{1j} \sim N(0, \tau_{00})$. Or, in short,

$$\beta_{qj} = \gamma_{q0} + u_{1j} \text{ for } q = 1,...,n. \tag{6-2-3.15}$$

Notice that u_{1j} is the same for all items and reflects the idea of unidimensionality of the IRT model, that all the items in a test are supposed to measure a single construct. Notice also that in the Rasch model it is not necessary to fix $\tau_{00} = 1$, though in IRT it is customary to do so. On the other hand, fixing some parameters is necessary for the 2-P IRT model in order for the model to be identified.

If we put the L-1 and L-2 models together, we get the combined model.

Combined Model:

$$\eta_{ij} = \sum_{q=1}^{n} \gamma_{q0} X_{qij} + u_{1j} \sum_{q=1}^{n} X_{qij} \,. \tag{6-2-3.16}$$

But since $\sum_{q=1}^{n} X_{qij} \equiv 1$ by definition of indicator variables, it reduces to

$$\eta_{ij} = \sum_{q=1}^{n} \gamma_{q0} X_{qij} + u_{1j} \,. \tag{6-2-3.17}$$

I now show that the model in Equation (6-2-3.17) is equivalent to the model formulated by Kamata (1998), and Cheong & Raudenbush (1999), that used an intercept model in order to fit the Rasch model to the existing HGLM setup. That is, at level-1, we let one item, for example, the first item, be the reference item, and create (n-1) dummy variables to represent differential item effects on response probability of student j.

L-1 structural Model:

$$\begin{aligned}\eta_{ij} &= \beta_{0j} + \beta_{1j}X_{2ij} + \beta_{2j}X_{3ij} + \cdots + \beta_{qj}X_{1ij} + \cdots + \beta_{(n-1)j}X_{nij} \\ &= \beta_{0j} + \sum_{q=1}^{n}\beta_{qj}X_{1ij}\end{aligned} \qquad (6\text{-}2\text{-}3.18)$$

where X_{qij} is the indicator variable and is 1 if $q = i$ and 0 if $q \neq i$ for all i. At level-2, only intercept β_{0j} varies.

L-2:

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + u_{0j} \\ \beta_{1j} &= \gamma_{10} \\ \beta_{2j} &= \gamma_{20} \\ &\cdots \\ \beta_{qj} &= \gamma_{q0} \\ &\cdots \\ \beta_{(n-1)i} &= \gamma_{(n-1)0}\end{aligned} \qquad (6\text{-}2\text{-}3.19)$$

where $u_{0j} \sim N(0, \tau_{00})$. The combined model can be written as,

Combined Model:

$$\eta_{ij} = \gamma_{00} + \sum_{q=1}^{n}\gamma_{q0}X_{qij} + u_{0j}. \qquad (6\text{-}2\text{-}3.20)$$

If we let $\delta_i \equiv -\{\gamma_{00} + \sum_{q=1}^{n} \gamma_{q0} X_{qij}\}$, $\theta_j \equiv u_{0j}$, then the combined model is exactly the same as the 1-P IRT Rasch model except that the constant $D \cong 1.7$, which adjusts it to the normal ogive model, was omitted. In IRT model terminology, δ_i represents *i*th item difficulty, θ_j is *j*th student ability. To interpret the combined model, we consider item 1 and item 2. For item 1, $\eta_{1j} = \gamma_{00} + u_{0j}$, and for item 2, $\eta_{1j} = \gamma_{00} + \gamma_{10} + u_{0j}$. Thus, the interpretation in words can be that the logit correct response probability of student *j* for item 1 is student's ability u_{0j} adjusted by the item difficulty γ_{00} (item easiness may reflect the exact meaning of the parameter) for item 1. And the logit correct response probability of student *j* for item 2 is student's ability u_{0j} adjusted by the item difficulty $(\gamma_{00} + \gamma_{10})$ for item 2. γ_{10} reflects the difference in terms of the item difficulty of item 2 compared to item 1. If item 2 is easier than item 1, then $\gamma_{10} > 0$ and thus $p_{2j} \geq p_{1j}$, where p_{1j} and p_{2j} are the correct response probabilities for item 1 and item 2 of student *j*.

It should be noted that the 1-P IRT model can be considered as a non-linear two-way ANOVA additive model with the factors, item difficulty and student ability.

2-P IRT model

Since using all *n* indicator variables for items without including the intercept is more natural, I'm going to adopt a no-intercept hierarchical model formulation to represent the following 2-P IRT model. The 2-P IRT model adds another parameter called item discrimination on each item. Item discrimination can be interpreted as a kind of

item's sharpness in distinguishing two students who have different abilities by the item. This concept suggests the interaction between item characteristics and student ability. That is, the effect of student ability on item response depends on an item characteristic which we call item discrimination even after controlling item difficulty. In other words, some items are sharper than others in reflecting student's ability. By this conceptualization, we formulate the level-1 structural model as

L-1:

$$\begin{aligned}\eta_{ij} &= \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \cdots + \beta_{qj}X_{1ij} + \cdots + \beta_{nj}X_{nij} \\ &= \sum_{q=1}^{n} \beta_{qj}X_{qij},\end{aligned} \tag{6-2-3.21}$$

or, in matrix notation,

$$\eta_{ij} = X_{ij}^{T}\beta_{j}, \tag{6-2-3.22}$$

where $X_{ij}^{T} = \left(X_{1ij},\ X_{2ij},\ \cdots,X_{qij},\ \cdots,X_{nij}\right)$. This notation is actually simpler that it seems: the *i*th element is 1 and the rest of them are 0, i.e.,

$X_{ij}^{T} = \left(0,\ 0,\ \cdots,1,\ \cdots,0\right)$. Note that this is the same as the level-1 model of the Rasch model. But the level-2 model is different.

L-2:

$$\begin{aligned}\beta_{1j} &= \gamma_{10} + \lambda_{1}v_{1j} \\ \beta_{2j} &= \gamma_{20} + \lambda_{2}v_{1j} \\ &\cdots \\ \beta_{qj} &= \gamma_{q0} + \lambda_{q}v_{1j} \\ &\cdots \\ \beta_{nj} &= \gamma_{n0} + \lambda_{n}v_{1j}\end{aligned} \tag{6-2-3.23}$$

where $v_{1j} \sim N(0,1)$. It should be noted that the variance of v_{1j} is now set to 1 for model identification. Notice the similarity and difference between the 2-P IRT model and the Rasch model in Equation (6-2-3.14). The same v_{1j} reflects that all items are measuring one kind of ability and the different coefficient λ implies that the sensitivity of each item to reflect ability v_{1j} is different; in other words, λ_q is the interaction effect between the membership of item q (X_{qij}) and person j's ability (v_{1j}). In matrix notation,

$$\beta_j = \gamma + \Lambda v_j, \; v_j \sim N(0,1) \qquad \text{(6-2-3.24)}$$

where $\beta_j = (\beta_{1j}, \beta_{2j}, \ldots, \beta_{nj})^T$, $\gamma = (\gamma_{10}, \gamma_{20}, \ldots, \gamma_{n0})^T$, $\Lambda = \left(\lambda_1, \; \lambda_2, \; \cdots, \; \lambda_n\right)^T$, $v_j = \left(v_{1j}\right)$.

If we combine the L-1 and L-2 models together, we get

$$\eta_{ij} = \{\sum_{q=1}^{n} \gamma_{qj} X_{qij}\} + \{\sum_{q=1}^{n} \lambda_q X_{qij}\} v_{1j}, \qquad \text{(6-2-3.25)}$$

and if we let $\mu_i \equiv \sum_{q=1}^{n} \gamma_{qj} X_{qij}$, $\lambda_i \equiv \sum_{q=1}^{n} \lambda_q X_{qij}$, and $\theta_j \equiv v_{0j}$ then the above equation can be written as

$$\eta_{ij} = \mu_i + \lambda_i \theta_j, \qquad \text{(6-2-3.26)}$$

where μ_i is the item intercept for item i, λ_i is the item discrimination for item i, and θ_j is the ability of student j. Using the item intercept as an item parameter is a way of representing the 2-P IRT model (See, for example, Bock & Aitkin, 1981), but if we impose a constraint on μ_i such as $\mu_i = -\lambda_i \delta_i$ for all i, then the model becomes

$$\eta_{ij} = \lambda_i (\theta_j - \delta_i) \qquad \text{(6-2-3.27)}$$

and this is the standard 2-P IRT model. We now write the models for all n items together. At level-1 we have

L-1:

$$\eta_j = A_j \beta_j \qquad (6\text{-}2\text{-}3.28)$$

where $\eta_j = (\eta_{1j},\ldots,\eta_{nj})^T$, $A_j = \begin{pmatrix} X_{1j}^T \\ \vdots \\ X_{nj}^T \end{pmatrix}$, and $\beta_j = (\beta_{1j},\ldots,\beta_{nj})^T$. Note that A_j is the $n \times n$ identity matrix, i.e., $A_j = \mathbf{I}_n$ if student i responded to all of the items, which is a case of balanced design. The level-2 model is written as Equation (6-2-3.24), and thus the combined model is,

Combined Model:

$$\eta_j = A_j \gamma + A_j \Lambda v_j,\ v_j \sim N(0, \Psi) \qquad (6\text{-}2\text{-}3.29)$$

where η_j is the $n \times 1$ outcome vector, γ is the $n \times 1$ item intercept vector, A_j is the $n \times n$ design matrix, Λ is the $n \times 1$ item discrimination vector (later, it will be $n \times M$ matrix if we use M multidimensional IRT model), v_j is the 1×1 scalar (later we extend it to a $M \times 1$ vector), and Ψ is a scalar variance of latent ability of student j (later, a $M \times M$ covariance matrix). For the above 3 items case, Equation (6-2-3.29) becomes

$$\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \gamma_{00} \\ \gamma_{10} \\ \gamma_{20} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \end{pmatrix} \left(v_{1j} \right). \qquad (6\text{-}2\text{-}3.30)$$

Note that the Rasch Model is a special case of the 2-P IRT model in that it lets $\Lambda = \mathbf{1}$, a vector of all 1's in Equation (6-2-3.29) or (6-2-3.30). Note that though the design matrix

A_j seems trivial ($A_j = \mathbf{I}_n$ for all i, which is true if all the students take the same set of items and if there is no missing observation), it can be different from student to student.

Now, notice the close similarity of Equation (6-2-3.29) to Equation (3-9), which is the HLM2F model developed in this dissertation. In fact, by letting $u_j = \Lambda v_j$ and denoting $D[u_j] \equiv \tau$ in Equation (6-2-3.29), we have $\tau = \Lambda\Psi\Lambda^T$. Thus, by formulating the IRT model by HGLM with a factor structure, we recognize that the standard 2-parameter IRT model is a non-linear version of the model Equation (3-9), which I call HGLM2F. This is the reason why I say the IRT model is a non-linear version of HLM2F. Instead of standardizing $\psi_{00} = 1$, which is the standard procedure of IRT programs, we can set $\lambda_0 = 1$ for model identification. Then, we can directly incorporate the algorithm of HLM2F into the micro iteration of HGLM, which consists of a doubly iterative procedure.

Multi-dimensional IRT model

The nice thing about this formulation is that we can easily expand it to a multidimensional model. That is, if we think that the test is measuring multiple abilities, then we change Equation (6-2-3.9) to

$$\begin{pmatrix} \eta_{1j} \\ \eta_{2j} \\ \eta_{3j} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} + \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \end{pmatrix} \begin{pmatrix} \theta_{1j} \\ \theta_{2j} \end{pmatrix}, \qquad (6\text{-}2\text{-}3.31)$$

and Equation (6-2-3.12) to

$$\begin{pmatrix} \theta_{1j} \\ \theta_{2j} \end{pmatrix} = \begin{pmatrix} \beta_{00} \\ \beta_{10} \end{pmatrix} + \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix}, \qquad (6\text{-}2\text{-}3.32)$$

where $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} \sim N_2(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{I}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})$. Note that the covariance matrix of $\begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix}$ was set to the identity matrix so that the model can be identified except for the rotational indeterminacy.

For a HGLM formulation with dummy variables, we just need to change Λ by adding another column from $\Lambda = \begin{pmatrix} \lambda_1 \\ \lambda_{12} \\ \vdots \\ \lambda_q \\ \vdots \\ \lambda_n \end{pmatrix}$ to $\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \vdots & \vdots \\ \lambda_{q1} & \lambda_{q2} \\ \vdots & \vdots \\ \lambda_{n1} & \lambda_{n2} \end{pmatrix}$, and v_j from $\begin{pmatrix} v_{1j} \end{pmatrix}$ to $\begin{pmatrix} v_{1j} \\ v_{2j} \end{pmatrix}$ in Equation (6-2-3.29).

By this formulation, we can perform the unidimensionality test by formulating the null hypothesis $H_0: \lambda_{i2} = 0$ for all i. Further, if we model the level-2 by some individual characteristics, then we can perform a statistical test for Differential Item Functioning (DIF) (Hambleton, Swaminathan, & Rogers, 1991).

For the longitudinal data, we can extend the IRT model to a latent ability growth model based on this formulation. In that case, the scaling issue should be carefully addressed in order for the latent ability growth to be meaningful.

6-2-4. Other Possibilities

Incorporating the factor model into a three level model is a natural extension of the two level model. In the three level case, there are two possibilities of using the factor

analysis model; one is in level-2 covariance matrix τ_π, and another is in level-3 covariance matrix τ_β.

Incorporating the factor model into a three level multivariate hierarchical model is also straightforward for conceptualization.

As shown in the previous IRT application section, applying the factor analysis model to hierarchical generalized linear model (HGLM), which has a non-linear outcome, seems to be promising, at least to measurement models in psychometrics.

Appendix. Derivation and Algorithm

The purpose of this appendix is to derive the Maximum Likelihood Estimators (MLE) via Fisher scoring algorithm and to give the detailed steps of algorithm for computation.

Model:

Subscript Model (Model for each group)

$$Y_j = A_{1j}\theta_1 + A_{2j}\Lambda\eta_j + r_j, (j = 1,2,...,J) \quad \text{(A-1)}$$

where

$$r_j \sim N(0,\sigma^2\mathbf{I}_{n_j}),$$

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

and r_j and η_j are independent for all j. Y_j is $n_j \times 1$, A_{1j} is $n_j \times F$, θ_1 is $F \times 1$, η_j is the $M \times 1$ factor score vector, r_j is a $n_j \times 1$; A_{2j} is a $n_j \times R$, Λ is $R \times M$, $\mathbf{I}_{n_j}$ is the $n_j \times n_j$ identity matrix, and Ψ is the positive definite $M \times M$ symmetric matrix.

By writing

$$e_j = A_{2j}\Lambda\eta_j + r_j \quad \text{(A-2)}$$

in Equation (A-1), we see the model (A-1) as general linear model,

$$Y_j = A_{1j}\theta_1 + e_j. \quad \text{(A-3)}$$

The variance-covariance matrix is

$$V_j \equiv Var(e_j) = A_{2j}\Lambda\Psi\Lambda^T A_{2j}^T + \sigma^2\mathbf{I}_{n_j}. \quad \text{(A-4)}$$

The log-likelihood:

The Log-Likelihood l is

$$l = -\frac{1}{2}[N\log(2\pi) + (N - JM)\log(\sigma^2) + J\log|\Psi| - \sum_{j=1}^{J}\log|C_j^{-1}| + \sum_{j=1}^{J}(e_j^T e_j - e_j^T A_{2j}\cdot\Lambda C_j^{-1}\Lambda^T\cdot A_{2j}^T e_j)/\sigma^2], \quad \text{(A-5)}$$

where $C_j^{-1} = (\Lambda^T A_{2j}^T A_{2j}\Lambda + \sigma^2\Psi^{-1})^{-1}$ that will be explained in Equation (A-16),

$e_j^T e_j$ is given by Equation (A-27), and

$A_{2j}^T e_j$ is given by Equation (A-28), to be shown later.

Computation:

In the following, the subscript for the A_2 will be omitted for simplicity of the formula and will be written as A, and all the subscript j will be omitted. If it is necessary for clarity, we will explicitly use the subscript.

Review of Standard 2-level HLM

Elements required for standard 2 level MLF, Fisher Scoring Algorithm are:

$E = \frac{dvec(\tau)}{d\phi^T}$, $F = \frac{d\sigma^2}{d\phi^T}$, and for $H = E[\frac{\partial^2 \log L(\theta_1, \tau, \sigma^2; \mathbf{y})}{\partial\phi^T\partial\phi}]$, $A^T V^{-1} A$, $A^T V^{-2} A$, $tr(V^{-2})$, and $e^T V^{-2} e$ since

$$H = -\frac{1}{2}\left(\frac{\partial vec(V)}{\partial\phi^T}\right)^T (V^{-1}\otimes V^{-1})\frac{\partial vec(V)}{\partial\phi^T} \quad \text{(A-6)}$$
$$= -\frac{1}{2}[E^T(A^T V^{-1} A\otimes A^T V^{-1} A)E + E^T vec(A^T V^{-2} A)F + F^T vec(A^T V^{-2} A)E + tr(V^{-2})F^T F];$$

for $S = \frac{\partial \log L(\theta_1, \tau, \sigma^2; \mathbf{y})}{\partial\phi}$, E, F, $A^T V^{-1} e$, $tr(V^{-1})$, and $e^T V^{-2} e$ are required since

$$S = \frac{1}{2}(\frac{\partial vec(V)}{\partial \phi^T})^T (V^{-1} \otimes V^{-1}) vec(ee^T - V)$$
$$= \frac{1}{2}[E^T vec(A^T V^{-1} ee^T V^{-1} A - A^T V^{-1} A) + F^T \{e^T V^{-2} e - tr(V^{-1})\}]. \quad \text{(A-7)}$$

Now for 2 level MLF factor model, we need the same things. But there are some differences, because in this case,

$$\tau = \Lambda \Psi \Lambda^T \quad \text{(A-8)}$$

τ is $R \times R$ and is not full and matrix which is **not invertible,**

Λ is $R \times M$, and Ψ is a $M \times M$ positive definite symmetric matrix.

(1) *Computation of E and F*

Definition 1. **Definition of ϕ vector**

$$\phi = (\{\lambda_{ij}\}, \{\psi_{hl}\}, \sigma^2)^T$$
$$= (\lambda^T, \psi^T, \sigma^2)^T, \quad \text{(A-9)}$$

is a $Q \times 1$ vector of unknown unique parameters in the variance covariance matrix V_j for all j, and it is composed of three elements, a $Q_1 \times 1$ vector λ, a $Q_2 \times 1$ vector ψ, a scalar σ^2.

(a) A $Q_1 \times 1$ vector λ consists of Q_1 unknown elements in the $R \times M$ factor loading matrix Λ ($Q_1 \leq RM$). Thus, some of the elements (Q_1) are unknown and to be estimated from the data, and the rest of the elements ($RM - Q_1$) are known. To construct λ, we align the unknown elements in the order of starting from (1,1) element of Λ and going down the column, if we find the unknown element, then put it in the λ vector. Then,

move on to next element in the same column of Λ. Once we finish searching in the first column, we move to the second column, and so on to the last Mth column of Λ.

Example

Ex 1.

For example, if $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix}$ then $\lambda = \begin{pmatrix} \lambda_{21} \\ \lambda_{42} \end{pmatrix}$.

Ex 2.

If $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & \lambda_{22} \\ 0 & 1 \\ \lambda_{41} & \lambda_{42} \end{pmatrix}$, then $\lambda = \begin{pmatrix} \lambda_{21} \\ \lambda_{41} \\ \lambda_{22} \\ \lambda_{42} \end{pmatrix}$.

Note that in exploratory factor analysis model, we let $\Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \\ \lambda_{31} & \lambda_{32} \\ \lambda_{41} & \lambda_{42} \end{pmatrix}$, all the elements in the Λ (factor loading matrix) are unknown. But, we are here thinking about confirmatory factor analysis in which we specify some of the elements as fixed number. In the first example, we fixed $\lambda_{11} = 1, \lambda_{31} = 0, \lambda_{41} = 0, \lambda_{21} = 0, \lambda_{22} = 0, \lambda_{32} = 1$, and in the second example, we fixed as $\lambda_{11} = 1, \lambda_{31} = 0,, \lambda_{21} = 0, \lambda_{32} = 1$. Thus, in our model, some of the elements in Λ are known and some of them are unknown. The number of known elements is $RM - Q_1$ and the number of unknown elements is Q_1.

(b) A $Q_2 \times 1$ vector ψ consists of Q_2 elements, where Q_2 be the number of unique elements in $M \times M$ variance-covariance matrix Ψ, and since Ψ is always symmetric,

the number of the unique elements Q_2 is $Q_2 = M(M+1)/2$. We align the

$Q_2 \times 1$ vector ψ as follows.

$$\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} & \cdots & \psi_{1M} \\ \psi_{21} & \psi_{22} & & \\ \vdots & & \ddots & \\ \psi_{M1} & & & \psi_{MM} \end{pmatrix} \Rightarrow \psi = (\psi_{11}, \psi_{21}, \cdots, \psi_{M1}, \psi_{22}, \psi_{23}, \cdots, \psi_{M2}, \psi_{33}, \cdots, \psi_{MM})^T .$$

This operation is often written as $\psi = vech(\Psi)$ (vector half).

Adding the number of unique elements in σ^2, which is clearly 1, we have in total,

$$Q = Q_1 + Q_2 + 1 = Q_1 + \frac{M(M+1)}{2} + 1$$

elements in the vector ϕ.

Note.

$Max(Q_1) = RM$ and $Q_2 = M(M+1)/2$. Thus, Q_2 is always exactly $M(M+1)/2$, the number of unique elements in the symmetric matrix Ψ, but Q_1 satisfies inequality restrictions $Q_1 \leq RM$ because we fix some of the elements in Λ, depending on our theory.

Definition 2. **Definition of E matrix**

We define E matrix as

$$E = \frac{dvec\tau}{d\phi^T} = (\frac{dvec\tau}{d\lambda^T}, \frac{dvec\tau}{d\psi^T}, \frac{dvec\tau}{d\sigma^{2T}}) = (E_\lambda, E_\psi, E_{\sigma^2}) . \quad \text{(A-10)}$$

The dimension of E is a $R^2 \times Q$ matrix, and E matrix consists of three parts, E_λ, E_ψ, and E_{σ^2}, which is always a $R^2 \times 1$ vector of 0 because τ does not depend on σ^2. Thus, E matrix always has a form

$$E = (E_\lambda, E_\psi, \mathbf{0}). \tag{A-11}$$

And, E_λ is a $R^2 \times Q_1$ matrix, E_ψ is a $R^2 \times Q_2$ matrix.

Now,

(a) For E_λ, we compute this column by column. The qth column of E_λ corresponds to $\frac{dvec\tau}{d\lambda_{ij}}$, where λ_{ij} is the qth element of λ (that is, the qth unknown element in Λ). The order is determined by counting for the same column down the rows and then move to the next column, go down the rows. Thus,

$$\begin{aligned} E_{\lambda_{ij}} &= \frac{\partial vec\tau}{\partial \lambda_{ij}} = \frac{\partial vec(\Lambda\Psi\Lambda^T)}{\partial \lambda_{ij}} = vec(\frac{\partial(\Lambda\Psi\Lambda^T)}{\partial \lambda_{ij}}) \\ &= vec(\frac{\partial \Lambda}{\partial \lambda_{ij}}\Psi\Lambda^T + \Lambda\Psi\frac{\partial \Lambda^T}{\partial \lambda_{ij}}) = vec(D_{\lambda_{ij}}\Psi\Lambda^T + \Lambda\Psi D_{\lambda_{ij}}^T), \end{aligned} \tag{A-12}$$

where $D_{\lambda_{ij}} = \frac{\partial \Lambda}{\partial \lambda_{ij}} = \{1$ for position (i,j) in Λ and 0 for other positions in $\Lambda\}$ for all i, j.

Example.

when $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix}$, $E_\lambda = (E_{\lambda_{21}}, E_{\lambda_{42}}) = (\frac{\partial vec\tau}{\partial \lambda_{21}}, \frac{\partial vec\tau}{\partial \lambda_{42}})$. Then,

$$E_{\lambda_{21}} = vec(D_{\lambda_{21}}\Psi\Lambda^T + \Lambda\Psi D_{\lambda_{21}}^T) = vec\{D_{\lambda_{21}}\Psi\Lambda^T + (D_{\lambda_{21}}\Psi\Lambda^T)^T\}.$$

$$D_{\lambda_{21}} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix},$$

$$D_{\lambda_{21}}\Psi\Lambda^T = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}\begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}\begin{pmatrix} 1 & \lambda_{21} & 0 & 0 \\ 0 & 0 & 1 & \lambda_{42} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \psi_{11} & \lambda_{21}\psi_{11} & \psi_{12} & \lambda_{42}\psi_{42} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \text{ and}$$

$$D_{\lambda_{21}}\Psi\Lambda^T + \Lambda\Psi D_{\lambda_{21}}^T = \begin{pmatrix} 0 & 0 & 0 & 0 \\ \psi_{11} & \lambda_{21}\psi_{11} & \psi_{12} & \lambda_{42}\psi_{42} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & \psi_{11} & 0 & 0 \\ 0 & \lambda_{21}\psi_{11} & 0 & 0 \\ 0 & \psi_{12} & 0 & 0 \\ 0 & \lambda_{42}\psi_{42} & 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0 & \psi_{11} & 0 & 0 \\ \psi_{11} & 2\lambda_{21}\psi_{11} & \psi_{12} & \lambda_{42}\psi_{42} \\ 0 & \psi_{12} & 0 & 0 \\ 0 & \lambda_{42}\psi_{42} & 0 & 0 \end{pmatrix}$$

Then taking vec, we obtain $E_{\lambda_{21}}$, which is a 16 × 1 column vector. Explicitly, it is

$E_{\lambda_{21}} = (0, \psi_{11}^{(0)}, 0, 0, \psi_{11}^{(0)}, 2\lambda_{21}\psi_{11}^{(0)}, \psi_{12}^{(0)}, \lambda_{42}^{(0)}\psi_{12}^{(0)}, 0, \psi_{12}^{(0)}, 0, 0, 0, \lambda_{42}^{(0)}\psi_{12}^{(0)}, 0, 0)^T$, where the superscript on the parameters means that we use the current values of estimates to compute E_{λ_1}. Similarly,

$E_{\lambda_{42}} = (0, 0, 0, \psi_{21}^{(0)}, 0, 0, 0, \lambda_{21}\psi_{21}^{(0)}, 0, 0, 0, \psi_{22}^{(0)}, \psi_{21}^{(0)}, \lambda_{21}^{(0)}\psi_{21}^{(0)}, \psi_{22}^{(0)}, 2\lambda_{42}^{(0)}\psi_{22}^{(0)})^T$.

(b) Next, for E_ψ, we compute it like the old τ in HLM2. That is, we can compute E_ψ one at a time by

$$E_\psi = (\Lambda \otimes \Lambda)D_\psi^*, \tag{A-13}$$

where

$$D_{\psi}^{*} = \frac{\partial vec\Psi}{\partial \psi^{T}},$$

where ψ is the part in the vector ϕ whose elements are the unique elements of Ψ, as defined in (A-9) in Definition 1.

Example. when $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix}$, $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$, and $\psi = \begin{pmatrix} \psi_{11}, & \psi_{21}, & \psi_{22} \end{pmatrix}^{T}$, then

$$D_{\psi}^{*} = \frac{\partial vec\Psi}{\partial \psi^{T}} = \frac{\partial \begin{pmatrix} \psi_{11} \\ \psi_{21} \\ \psi_{12} \\ \psi_{22} \end{pmatrix}}{\partial \begin{pmatrix} \psi_{11} & \psi_{21} & \psi_{22} \end{pmatrix}} = \left(\frac{\partial vec\psi}{\partial \psi_{11}}, \frac{\partial vec\psi}{\partial \psi_{21}}, \frac{\partial vec\psi}{\partial \psi_{22}} \right)$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, by Equation A-13, we get the 16×3 matrix E_{ψ} by

$$E_{\psi} = (\Lambda \otimes \Lambda) D_{\psi}^{*} = \left\{ \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix} \right\} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

$$= (\Lambda \otimes \Lambda)(D_{\psi_{11}}^{*}, D_{\psi_{21}}^{*}, D_{\psi_{22}}^{*})$$

$$= (\Lambda \otimes \Lambda)(vecD_{\psi_{11}}, vecD_{\psi_{21}}, vecD_{\psi_{22}})$$

$$= \left((\Lambda \otimes \Lambda) vecD_{\psi_{11}}, \quad (\Lambda \otimes \Lambda) vecD_{\psi_{21}}, \quad (\Lambda \otimes \Lambda) vecD_{\psi_{22}} \right)$$

$$\equiv \left(E_{\psi_{11}} \quad E_{\psi_{21}} \quad E_{\psi_{22}} \right)$$

$$= \begin{pmatrix} 1 & 0 & 0 \\ \lambda_{21} & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{42} & 0 \\ \lambda_{21} & 0 & 0 \\ \lambda_{21}^2 & 0 & 0 \\ 0 & \lambda_{21} & 0 \\ 0 & \lambda_{21}\lambda_{42} & 0 \\ 0 & 1 & 0 \\ 0 & \lambda_{21} & 0 \\ 0 & 0 & 1 \\ 0 & 0 & \lambda_{42} \\ 0 & \lambda_{42} & 0 \\ 0 & \lambda_{21}\lambda_{42} & 0 \\ 0 & 0 & \lambda_{42} \\ 0 & 0 & \lambda_{42}^2 \end{pmatrix}.$$

Definition 3. **Definition of F vector**

A $1 \times Q$ row vector F is defined by

$$F = \frac{\partial \sigma^2}{\partial \phi^T}. \tag{A-14}$$

Example.

when $\Lambda = \begin{pmatrix} 1 & 0 \\ \lambda_{21} & 0 \\ 0 & 1 \\ 0 & \lambda_{42} \end{pmatrix}$, $\Psi = \begin{pmatrix} \psi_{11} & \psi_{12} \\ \psi_{21} & \psi_{22} \end{pmatrix}$, $\phi = (\lambda_{21}, \lambda_{42}, \psi_{11}, \psi_{21}, \psi_{22}, \sigma^2)^T$, a 6×1 column

vector and $Q = Q_1 + Q_2 + 1 = 2 + \frac{M(M+1)}{2} + 1 = 2 + \frac{2(2+1)}{2} + 1 = 2 + 3 + 1 = 6$. Then,

$F = (0, \ 0, \ 0, \ 0, \ 0, \ 1)$.

(2) *Computation of H (Expected Hessian matrix) and S (Score vector)*

Note that H is a $Q \times Q$ matrix, and S is a $Q \times 1$ vector.

(a) Computation of H (expected Hessian matrix).

$$H^{(i)} = [E\{\frac{\partial^2 \log L(Y; \gamma, \Lambda, \Psi, \sigma^2)}{\partial \phi \partial \phi^T}\}]_{\phi = \phi^{(i)}} = \sum_{j=1}^{J} H_j^{(i)},$$

where

$$\begin{aligned} H_j^{(i)} &= -\frac{1}{2}\{\left(\frac{\partial vec(V_j)}{\partial \phi^T}\right)^T \left(V_j^{-1} \otimes V_j^{-1}\right)\left(\frac{\partial vec(V_j)}{\partial \phi^T}\right)\}|_{\phi = \phi^{(i)}} \\ &= -\frac{1}{2}\{E^T(A_{2j}^T V_j^{-1} A_{2j} \otimes A_{2j}^T V_j^{-1} A_{2j})E + E^T vec(A_{2j}^T V_j^{-2} A_{2j})F \\ &\quad + F^T[vec(A_{2j}^T V_j^{-2} A_{2j})]^T E + tr(V_j^{-2}) \cdot F^T F\}|_{\phi = \phi^{(i)}}, \end{aligned} \qquad \text{(A-15)}$$

where

$$E = \frac{\partial vec(\tau)}{\partial \phi^T}, \ F = \frac{\partial \sigma^2}{\partial \phi^T}.$$

To compute $H_j^{(i)}$, we need to know $A_{2j}^T V_j^{-1} A_{2j}$, $A_{2j}^T V_j^{-2} A_{2j}$, and $tr(V_j^{-2})$.

These quantities are function of Weighted Sufficient Statistics (WSS), which are

WSS = $\begin{bmatrix} A_{1j}^T V_j^{-1} A_{1j} & A_{2j}^T V_j^{-1} A_{2j} \\ A_{1j}^T V_j^{-1} A_{2j} & A_{2j}^T V_j^{-1} Y_j \\ A_{1j}^T V_j^{-1} Y_j & \end{bmatrix}$, for all j, the WSS is in the functions of Sufficient Statistics (SS), SS = $\begin{bmatrix} A_{1j}^T A_{1j} & A_{2j}^T A_{2j} \\ A_{1j}^T A_{2j} & A_{2j}^T Y_j \\ A_{1j}^T Y_j & \end{bmatrix}$, for all j, and $\Lambda C_j^{-1} \Lambda^T$ for all j. Therefore, we first show the computational formulae of $\Lambda C_j^{-1} \Lambda^T$ and the WSS, and then we show the computational formulae specific to $H_j^{(i)}$.

Formula of $\Lambda C_j^{-1} \Lambda^T$:

We first compute a $M \times M$ matrix

$$C_j^{-1} = (\Lambda^T A_{2j}^T A_{2j} \Lambda + \sigma^2 \Psi^{-1})^{-1}, j = 1,2,...,J, \qquad \text{(A-16)}$$

and then compute a $R \times R$ matrix $\Lambda C_j^{-1} \Lambda^T$ for all j. Then, we compute the WSS for the current $\Lambda C_j^{-1} \Lambda^T$.

Computation of WSS:

As we can see in the following, WSS is a functions of SS and $\Lambda C_j^{-1} \Lambda^T$.

- $A_{1j}^T V_j^{-1} A_{1j} = \sigma^{-2}(A_{1j}^T A_{1j} - A_{1j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T A_{1j})$; (A-17)
- $A_{1j}^T V_j^{-1} A_{2j} = \sigma^{-2}(A_{1j}^T A_{2j} - A_{1j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T A_{2j})$; (A-18)
- $A_{1j}^T V_j^{-1} Y_j = \sigma^{-2}(A_{1j}^T Y_j - A_{1j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T Y_j)$; (A-19)
- $A_{2j}^T V_j^{-1} A_{2j} = \sigma^{-2}(A_{2j}^T A_{2j} - A_{2j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T A_{2j})$; (A-20)
- $A_{2j}^T V_j^{-1} Y_j = \sigma^{-2}(A_{2j}^T Y_j - A_{2j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T Y_j)$. (A-21)

Computational formulae of $A_{2j}^T V_j^{-1} A_{2j}$, $A_{2j}^T V_j^{-2} A_{2j}$, and $tr(V_j^{-2})$ are as follows.

- $A_{2j}^T V_j^{-1} A_{2j}$ is computed by the formula (A-20).
- $A_{2j}^T V_j^{-2} A_{2j} = \sigma^{-2}(A_{2j}^T V_j^{-1} A_{2j} - A_{2j}^T V_j^{-1} A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T A_{2j})$ (A-22)
- $tr(V_j^{-2}) = (n_j - M)\sigma^{-4} + tr\{(C_j^{-1}\Psi^{-1})^2\}$. (A-23)

(b) Computation of S (Score vector).

$$S^{(i)} = \frac{\partial \log L(Y;\gamma,\Lambda,\tau,\sigma^2)}{\partial \phi}\bigg|_{\phi=\phi^{(i)}} = \sum_{j=1}^{J} S_j^{(i)},$$

where

$$\begin{aligned} S_j^{(i)} &= \frac{1}{2}[\{\frac{\partial vec(V_j)}{\partial \phi}\}^T \{(V_j^{-1} \otimes V_j^{-1}) vec(e_j e_j^T - V_j)\}]\Big|_{\phi=\phi^{(i)}} \\ &= \frac{1}{2}[E^T vec(A_{2j}^T V_j^{-1} e_j e_j^T V_j^{-1} A_{2j} - A_{2j}^T V_j^{-1} A_{2j}) + F^T \{e_j^T V_j^{-2} e_j - tr(V_j^{-1})\}]|_{\phi=\phi^{(i)}}. \end{aligned} \quad \text{(A-24)}$$

To compute $S_j^{(i)}$, we need to know $A_{2j}^T V_j^{-1} e_j$, $A_{2j}^T V_j^{-1} A_{2j}$, $e_j^T V_j^{-2} e_j$, and $tr(V_j^{-1})$.

Thus we need to know computational formulae of $A_{2j}^T V_j^{-1} e_j$, $A_{2j}^T V_j^{-1} A_{2j}$, $e_j^T V_j^{-2} e_j$, and $tr(V_j^{-1})$. To compute $A_{2j}^T V_j^{-1} e_j$, $A_{2j}^T V_j^{-1} A_{2j}$, $e_j^T V_j^{-2} e_j$, and $tr(V_j^{-1})$, we need to have evaluated e_j, $\theta_1^{(i)}$, $e_j^T e_j$, and $A_{2j}^T e_j$. Those quantities are computed as in the following.

We first note that

$$e_j = Y_j - A_{1j}\theta_1^{(i)}, \quad \text{(A-25)}$$

where $\theta_1^{(i)}$ is the current estimate of θ_1, and is computed by

$$\theta_1^{(i)} = (\sum_{j=1}^{J} A_{1j}^T V_j^{-1} A_{1j})^{-1} \sum_{j=1}^{J} A_{1j}^T V_j^{-1} Y_j. \quad \text{(A-26)}$$

Note that this is a function of WSS, i.e., $\theta_1^{(i)} = f(\text{WSS})$. Next we compute $e_j^T e_j$ and

$A_{2j}^T e_j$ by

$$e_j^T e_j = Y_j^T Y_j - 2\theta_1^{(i)T} A_{1j}^T Y_j + \theta_1^{(i)T} A_{1j}^T A_{1j} \theta_1^{(i)}; \quad \text{(A-27)}$$

$$A_{2j}^T e_j = A_{2j}^T Y_j - A_{2j}^T A_{2j} \theta_1^{(i)}, \text{ for all } j, \quad \text{(A-28)}$$

which are also functions of SS and WSS, i.e., $e_j^T e_j = f(\text{SS, WSS})$ and

$A_{2j}^T e_j = f(\text{SS, WSS})$. Then, using these values, we can compute the quantities necessary

to compute $S_j^{(i)}$.

$$A_{2j}^T V_j^{-1} e_j = A_{2j}^T V_j^{-1} Y_j - A_{2j}^T V_j^{-1} A_{2j} \theta_1^{(i)}; \quad \text{(A-29)}$$

$A_{2j}^T V_j^{-1} A_{2j}$ was already given in formula (A-18) in the computation of $H_j^{(i)}$;

$$e_j^T V_j^{-2} e_j = \sigma^{-4}(e_j^T e_j - e_j^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T e_j + e_j^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T A_{2j} \cdot \Lambda C_j^{-1} \Lambda^T \cdot A_{2j}^T e_j) \quad \text{(A-30)}$$

$$tr(V_j^{-1}) = (n_j - M)\sigma^{-2} + tr(C_j^{-1} \Psi^{-1}). \quad \text{(A-31)}$$

<u>Overall Steps via Fisher Scoring Algorithm</u>:

Step 1. Compute and Store the Sufficient Statistics (SS), i.e.,

$$\text{SS} = \begin{bmatrix} A_{1j}^T A_{1j} & A_{2j}^T A_{2j} \\ A_{1j}^T A_{2j} & A_{2j}^T Y_j \\ A_{1j}^T Y_j & \end{bmatrix} \quad \text{(A-32)}$$

for all j.

Step 2. Compute Statistics values $\theta_1^{(0)}$, $\phi^{(0)} = (\lambda^{(0)T}, \psi^{(0)T}, \sigma^{2(0)})^T$ by

$$\theta_1^{(0)} = (\sum_{j=1}^{J} A_{1j}^T A_{1j})^{-1} (\sum_{j=1}^{J} A_{1j}^T Y_j), \tag{A-33}$$

and compute the initial values of $\lambda^{(0)}$ and $\psi^{(0)}$ by the method which is provided in the next section..

Note: $\sigma^{2(0)}$ is computed by first computing

$$\theta_{2j}^{(0)} = (A_{2j}^T A_{2j})^{-1} (A_{2j}^T Y_j - A_{2j}^T A_{1j} \theta_1^{(0)}), \tag{A-34}$$

and then

$$\sigma^{2(0)} = \frac{1}{N - JR} \sum_{j=1}^{J} (Y_j - A_{1j}\theta_1^{(0)} - A_{2j}\theta_{2j}^{(0)})^T (Y_j - A_{1j}\theta_1^{(0)} - A_{2j}\theta_{2j}^{(0)}). \tag{A-35}$$

Step 3. Compute a $M \times M$ matrix $C_j^{-1} = (\Lambda^T A_{2j}^T A_{2j} \Lambda + \sigma^2 \Psi^{-1})^{-1}$, $j = 1,2,...,J$, by constructing Λ, Ψ from $\phi^{(0)}$, and then compute $\Lambda C_j^{-1} \Lambda^T$ for all j.

Note: We construct Λ, Ψ from $\phi^{(0)}$, and use them to compute C_j^{-1} with $\sigma^{2(0)}$.

Step 4. Compute the Weighted Sufficient Statistics (WSS), i.e.,

$$\text{WSS} = \begin{bmatrix} A_{1j}^T V_j^{-1} A_{1j} & A_{2j}^T V_j^{-1} A_{2j} \\ A_{1j}^T V_j^{-1} A_{2j} & A_{2j}^T V_j^{-1} Y_j \\ A_{1j}^T V_j^{-1} Y_j & \end{bmatrix} \tag{A-36}$$

for all j.

Step 5. Compute the matrix E and the vector F.

Note 1: F doesn't change, though E changes for each Fisher Scoring iteration.

Note 2: $E^{(0)} = f(\phi^{(0)})$. That is, $E^{(0)}$ depends on $\phi^{(0)}$ and $E^{(i)}$ will change as $\phi^{(i)}$ changes.

Step 6. Compute H, S by the formulae in Equation (A-15) and (A-24) along with computational formulae of the components.

Note: $H, S = f(E,\ \text{WSS})$.

Step 7. Compute the new $\phi^{(1)}$ by Fisher Scoring Algorithm, .i.e.,

$$\phi^{(1)} = \phi^{(0)} - H^{-1}S. \tag{A-37}$$

Note: $\phi^{(1)} = f(H^{-1}S)$.

Step 8. Compute the new

$$\theta_1^{(1)} = (\sum_{j=1}^{J} A_{1j}^T V_j^{-1} A_{1j})^{-1} \sum_{j=1}^{J} A_{1j}^T V_j^{-1} Y_j \tag{A-38}$$

based on the new $\phi^{(1)}$.

Note: $\theta_1^{(1)} = f(\text{WSS})$

Step 9. Compute the Log-Likelihood l.

The Log-Likelihood l is computed by

$$l = -\frac{1}{2}[N\log(2\pi) + (N - JM)\log(\sigma^2) + J\log|\Psi| - \sum_{j=1}^{J}\log|C_j^{-1}| + \sum_{j=1}^{J}(e_j^T e_j - e_j^T A_{2j} \cdot \Lambda C_j^{-1}\Lambda^T \cdot A_{2j}^T e_j)/\sigma^2],$$

as in Equation (A-5), where

$e_j^T e_j$ is computed by Equation (A-27), and $A_{2j}^T e_j$ is computed by Equation (A-28).

Step 10. Go back to Step 3.

Note: We use the same cutoff criterion as current HLM2, i.e., change in log-likelihood, to get out of the loop.

Starting Values:

To start the Fisher Scoring, we use the following quantities.

$$\theta_1^{(0)} = (\sum_{j=1}^{J} A_{1j}^T A_{1j})^{-1} (\sum_{j=1}^{J} A_{1j}^T Y_j); \quad \text{(A-39)}$$

$$\theta_{2j}^{(0)} = (A_{2j}^T A_{2j})^{-1} (A_{2j}^T Y_j - A_{2j}^T A_{1j} \theta_1^{(0)}); \quad \text{(A-40)}$$

$$\sigma^{2(0)} = \frac{1}{N - JR} \sum_{j=1}^{J} (Y_j - A_{1j}\theta_1^{(0)} - A_{2j}\theta_{2j}^{(0)})^T (Y_j - A_{1j}\theta_1^{(0)} - A_{2j}\theta_{2j}^{(0)}); \quad \text{(A-41)}$$

$$\bar{V} = \frac{\sigma^{2(0)}}{J} \sum_{j=1}^{J} (A_{2j}^T A_{2j})^{-1}; \quad \text{(A-42)}$$

$$\bar{D} = \frac{1}{J} \sum_{j=1}^{J} \theta_{2j}^{(0)} \theta_{2j}^{(0)T}; \quad \text{(A-43)}$$

$$\bar{\tau} = \bar{D} - \bar{V}; \quad \text{(A-44)}$$

$\bar{\tau}$ is the method of moments estimate in a balanced design. At this point, we check the positive semi-definitness of the $\bar{\tau}$ matrix. If the $\bar{\tau}$ matrix is not positive semi-definite (p.s.d.), we fix it up in the following way:

1) If $\bar{\tau}_{ii} \leq 0$, set $\bar{\tau}_{ii}$ to $.1\,\bar{D}_{ii}$.

2) If $|\bar{\tau}_{ij}| \geq \sqrt{\bar{\tau}_{ii}\bar{\tau}_{jj}}$, let $\bar{\tau}_{ij} = .9\sqrt{\bar{\tau}_{ii}\bar{\tau}_{jj}}$ (use the sign of the original $\bar{\tau}_{ij}$); this reduction is 10 percent from the original value.

3) Now check if $\bar{\tau}$ is p.s.d. If yes, go to the next step. If no, go back to step 2), and let $\bar{\tau}_{ij} = .8\sqrt{\bar{\tau}_{ii}\bar{\tau}_{jj}}$, another 10 percent reduction from the original value. Check if $\bar{\tau}$ is p.s.d. again. Keep repeating this process. If after five times of this repetition and if $\bar{\tau}$ is still not p.s.d., then set all of the off-diagonal elements to 0.

With the above procedure, we obtain a p.s.d. $\bar{\tau}$, an estimate of τ. Using this $\bar{\tau}$, we first obtain the starting values for Λ. We start with $\Psi = \mathbf{I}_M$. Then, since $\tau = \Lambda\Psi\Lambda^T$, we have $\tau = \Lambda\Lambda^T$. Since $\bar{\tau}$ is p.s.d., it can be decomposed into $\bar{\tau} = C\Delta C^T$, where Δ is an $R \times R$ matrix in which eigenvalues are on the diagonal, and C is an $R \times R$ matrix that the corresponding eigenvectors are in columns. We approximate C and Δ by taking the first M largest eigenvalues and the corresponding eigenvectors. We denote them as Δ^* and C^*. Note that Δ^* is a $M \times M$ matrix and C^* is $R \times M$ matrix. Then an equality holds approximately $\bar{\tau} = \Lambda^*\Lambda^{*T} \cong C^*\Delta^*C^{*T} = (C^*\Delta^{*\frac{1}{2}})(\Delta^{*\frac{1}{2}}C^{*T})$. Thus we obtain $\Lambda^* = C^*\Delta^{*\frac{1}{2}}$. Now we use the information for Λ specification. For clarity, we use an example.

Suppose that we specified the Λ as $\Lambda = \begin{pmatrix} 1 & 0 & \lambda_5 \\ 0 & 1 & \lambda_6 \\ 0 & \lambda_3 & 1 \\ \lambda_1 & \lambda_4 & 0 \\ \lambda_2 & 0 & 0 \end{pmatrix}$ and suppose we obtained

$\Lambda^* = \begin{pmatrix} \lambda_{11} & \lambda_{12} & \lambda_{13} \\ \lambda_{21} & \lambda_{22} & \lambda_{23} \\ \lambda_{31} & \lambda_{32} & \lambda_{33} \\ \lambda_{41} & \lambda_{42} & \lambda_{43} \\ \lambda_{51} & \lambda_{52} & \lambda_{53} \end{pmatrix}$ from the previous decomposition. Then, we let the

$$\hat{\Lambda}^{(0)} = \begin{pmatrix} 1 & 0 & \lambda_{13}/\lambda_{33} \\ 0 & 1 & \lambda_{23}/\lambda_{33} \\ 0 & \lambda_{32}/\lambda_{22} & 1 \\ \lambda_{41}/\lambda_{11} & \lambda_{42}/\lambda_{22} & 0 \\ \lambda_{51}/\lambda_{11} & 0 & 0 \end{pmatrix} \qquad \text{(A-46)}$$

and use it as the starting values for the estimate of Λ. In case the pivotal estimates in $\hat{\Lambda}^{(0)}$ are close to zero (in this example, either λ_{11} or λ_{22} or λ_{33}), then we use the initial estimate in $\Lambda^{\bullet}$ as it is. Note that we need to check if $rank(\hat{\Lambda}^{(0)}) = M$.

Next, we compute the starting value for the estimate of Ψ based on $\hat{\Lambda}^{(0)}$ and $\bar{\tau}$. Since $\tau = \Lambda\Psi\Lambda^T$ and $rank(\Lambda) = M$, we can obtain Ψ by $\Psi = (\Lambda^T\Lambda)^{-1}(\Lambda^T\tau\Lambda)(\Lambda^T\Lambda)^{-1}$ because $rank(\Lambda^T\Lambda) = M$ and thus it is invertible. Therefore, we let the starting value for the estimate of Ψ be

$$\hat{\Psi}^{(0)} = (\hat{\Lambda}^{(0)T}\hat{\Lambda}^{(0)})^{-1}(\hat{\Lambda}^{(0)T}\bar{\tau}\hat{\Lambda}^{(0)})(\hat{\Lambda}^{(0)T}\hat{\Lambda}^{(0)})^{-1}. \qquad \text{(A-47)}$$

Note that we need to check whether $\hat{\Psi}^{(0)}$ is positive semi-definite before we use it as the starting value.

We leave an option for the users that they can specify the starting values for $\hat{\Lambda}^{(0)}$ and $\hat{\Psi}^{(0)}$.

Endnote:

Derivation of log-likelihood formula in Equation (A-5):

In general, linear model with normal distribution as in Equation (A-3), the log-likelihood is given by

$$l = -\frac{N}{2}\log(2\pi) + \frac{1}{2}\sum_{j=1}^{J}\log|V_j^{-1}| - \frac{1}{2}\sum_{j=1}^{J}e_j^T V_j^{-1} e_j, \qquad \text{(B-1)}$$

where V_j^{-1} is defined in Equation (A-4) and e_j is defined in Equation (A-2). By Smith's Theorem 2, $(A\Omega A^T + \Psi)^{-1} = \Psi^{-1} - \Psi^{-1}A(A^T\Psi^{-1}A + \Omega^{-1})^{-1}A^T\Psi^{-1}$,

we have

$$V_j^{-1} = \sigma^{-2}(\mathbf{I}_{n_j} - A_{2j}\Lambda C_j^{-1}\Lambda^T A_{2j}^T),$$

where

$$C_j^{-1} = (\Lambda^T A_{2j}^T A_{2j}\Lambda + \sigma^2\Psi^{-1})^{-1}.$$

Now,

$$\log|V_j^{-1}| = \log|\sigma^{-2}(\mathbf{I}_{n_j} - A_{2j}\Lambda C_j^{-1}\Lambda^T A_{2j}^T)| = \log|\sigma^{-2}\mathbf{I}_{n_j}| + \log|\mathbf{I}_{n_j} - A_{2j}\Lambda C_j^{-1}\Lambda^T A_{2j}^T|$$

$$= -n_j \log(\sigma^2) + \log\begin{vmatrix} C_j & \Lambda^T A_{2j}^T \\ A_{2j}\Lambda & \mathbf{I}_{n_a} \end{vmatrix} |C_j^{-1}|$$

$$\left(\text{since}\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A||D - CA^{-1}B| (= |D||A - BD^{-1}C|), \quad |D - CA^{-1}B| = \begin{vmatrix} A & B \\ C & D \end{vmatrix} |A^{-1}|.\right)$$

$$= -n_j \log(\sigma^2) + \log\begin{vmatrix} C_j & \Lambda^T A_{2j}^T \\ A_{2j}\Lambda & \mathbf{I}_{n_a} \end{vmatrix} + \log|C_j^{-1}|$$

$$\left(\text{Again, apply}\begin{vmatrix} A & B \\ C & D \end{vmatrix} = (|A||D - CA^{-1}B|) = |D||A - BD^{-1}C|.\right)$$

$$= -n_j \log(\sigma^2) + \log|\mathbf{I}_{n_j}|\left|C_j - \Lambda^T A_{2j}^T \mathbf{I}_{n_j} A_{2j}\Lambda\right| + \log|C_j^{-1}|$$

$$= -n_j \log(\sigma^2) + \log(1) + \log\left|C_j - \Lambda^T A_{2j}^T A_{2j}\Lambda\right| + \log|C_j^{-1}|$$

$$= -n_j \log(\sigma^2) + \log\left|C_j - \Lambda^T A_{2j}^T A_{2j}\Lambda\right| + \log|C_j^{-1}|$$

$$= -n_j \log(\sigma^2) + \log|\sigma^2\Psi^{-1}| + \log|C_j^{-1}| \qquad (C_j^{-1} = (\Lambda^T A_{2j}^T A_{2j}\Lambda + \sigma^2\Psi^{-1})^{-1})$$

$$= -n_j \log(\sigma^2) + M\log(\sigma^2) + \log|\Psi^{-1}| + \log|C_j^{-1}| \qquad (|\Psi^{-1}| = |\Psi|^{-1})$$

$$= -\{(n_j - M)\log(\sigma^2) + \log|\Psi| - \log|C_j^{-1}|\}. \qquad \text{(B-2)}$$

Thus,

$$\sum_{j=1}^{J} \log|V_j^{-1}| = -\{(N - JM)\log(\sigma^2) + J\log|\Psi| - \sum_{j=1}^{J} \log|C_j^{-1}|\}. \tag{B-3}$$

And

$$\begin{aligned} e_j^T V_j^{-1} e_j &= e_j^T (\mathbf{I}_{n_j} - A_{2j}\Lambda C_j^{-1}\Lambda^T A_{2j}^T) e_j / \sigma^2 \\ &= (e_j^T e_j - e_j^T A_{2j} \cdot \Lambda C_j^{-1}\Lambda^T \cdot A_{2j}^T e_j) / \sigma^2, \end{aligned} \tag{B-4}$$

where $e_j^T e_j$ is computed by Equation (A-27), and $A_{2j}^T e_j$ is computed by Equation (A-28).

Thus,

$$\sum_{j=1}^{J} e_j^T V_j^{-1} e_j = \sum_{j=1}^{J} (e_j^T e_j - e_j^T A_{2j} \cdot \Lambda C_j^{-1}\Lambda^T \cdot A_{2j}^T e_j) / \sigma^2 \tag{B-5}$$

Therefore, plugging Equations (B-3) and (B-5) into Equation (B-1), we obtain the log-likelihood

$$l = -\frac{1}{2}[N\log(2\pi) + (N - JM)\log(\sigma^2) + J\log|\Psi| - \sum_{j=1}^{J} \log|C_j^{-1}| + \sum_{j=1}^{J} (e_j^T e_j - e_j^T A_{2j} \cdot \Lambda C_j^{-1}\Lambda^T \cdot A_{2j}^T e_j) / \sigma^2],$$

which is Equation (A-5). Notice that both of them are functions of WSS and $\Lambda C_j^{-1}\Lambda^T$.

References

Achenbach, T. M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington: University of Vermont Department of Psychiatry.

Adams, R. J., Wilson, M., and Wu, M. (1997). Multilevel Item response Models: An Approach to Errors in Variables Regression. *Journal of Educational and Behavioral Statistics, Spring 1997, Vol. 22, No. 1*, 47-76.

Aitkin, M., Anderson, D., & Hinde, J. (1981). Statistical Modeling of Data on Teaching Style. *Journal of Royal Statistical Society A, 144, Part 4,* 419-461.

Aitkin, M., & Longford, N. (1986). Statistical Modeling Issues in School Effectiveness Studies. *Journal of Royal Statistical Society A, 149, Part 1, 1*-43.

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*, second edition. New York, John Wiley & Sons.

Arbuckle, J. L. (1995). *Amos user's guide*. Chicago: SmallWaters.

Bartholomew, D. J., & Knott, M. (1999). *Latent Variable Models and Factor Analysis*, second edition. Kendall's Library of Statistics 7. London, Edward Arnold.

Becker, B. J. (1988). Synthesizing Standard Mean-change measures. *British Journal of Mathematical and Statistical Psychology 41,* 257-378.

Becker, B. J. (1990). Coaching for the Scholastic Aptitude Test: Further Synthesis and Appraisal. *Review of Educational Research, vol. 60, No. 3*, 373-417.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS for Windows User's Guide*. Encino, CA: Multivariate Software, Inc.

Bock, R. D. (1988). *Multilevel Analysis of Educational Data.* London: Academic Press.

Bock, R. D., & Aitkin, M. (1981). Marginal Maximum Likelihood Estimation of Item Parameters: Application of EM algorithm. *Psychometrika, 46,* 443-459.

Bollen, K. A. (1989). *Structural equations with latent variables.* New York,: Wiley.

Bollen, K. A. and Jöreskog, K. G. (1985). Uniqueness does not imply identification: A note on confirmatory factor analysis. *Sociological Methods and Research, 14,* 155-163.

Bryk, A. S., & Frank, K. (1991). The specialization of teachers' work: An initial exploratuion. In S. W. Raudenbush & J. D. Willms (Eds.), *Schools, classrooms, and pupils: International studies of schooling from a multilevel perspective* (pp. 185-204). Orlando, FL: Academic Press.

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

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *Hierarchical Linear and Nonlinear Modeling with the HLM/2L and HLM/3L.* Chicago: Scientific Software International. Inc.

Catchpole, E. A., & Morgan, B. J. T. (1994). Boundary estimation in ring recovery models. *Journal of Royal Statistical Society, Series B, 49,* 95-101.

Cheong, Y. F., & Raudenbush, S. W. (1999) *Measurement and Structural Models for Children's Problem Behaviors* (under review).

Chou, C., Bentler, P. M., & Pentz, M. A. (1998). Comparisons of Two Statistical Approaches to Study Growth Curves: The Multilevel Model and the Latent Curve Analysis. *Structural Equation Modeling, 5(3)*, 247-246.

Coleman, J. S., Hoffer, T., Kilgore, S B. (1982). *High school achievement: Public, Catholic and other schools compared.* New York: Basic Books.

Cronbach, L. J. (1951). Coefficients alpha and the internal structure of tests. *Psychometrika, 16*, 297-334.

Davis, W. R. (1993). The FCI rule of identification for confirmatory factor analysis: A general sufficient condition. *Sociological Methods and Research, 21*, 403-437.

De Leeuw, J., & Kreft, I. (1986). Random Coefficients Models for Multilevel Analysis. *Journal of Educational Statistics, 11*(1), 57-85.

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53*, 1-15.

Gan, L. & Jiang, J. (1999). A Test for Global Maximum. *Journal of American Statistical Association, September, 1999, Vol. 94, No. 447*, 847-854.

Garthwaite, P. H., Jolliffe, I. T., & Jones, B. (1995). *Statistical Inference*. London: Prentice Hall.

Gleser, L. J. and Olkin, I. (1994). Stochastically Dependent Effect Sizes. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis*, (pp. 339-355). New York: Russell Sage Foundation.

Goldstein, H. G. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. *Biometrika, 73*(1), 43-56.

Goldstein, H. G. (1995). *Multilevel Statistical Models*, Second Edition. Kendall's Library of Statistics 3. London: Edward Arnold.

Goldstein, H., and McDonald, R. P. (1988). A general model for the analysis of multilevel data. *Psychometrika, 53*. 435-67.

Goldstein, H., Rashbash, J., Plewis, I., Draper, D., Browne, W., Yan, M., Woodhouse, G., and Healy, M. (1998). *A User's Guide to MLWin*. London: University of London, Institute of Education.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, California: Sage Publication.

Hedeker, D., & Gibbons, R. D. (1996). MIXOR: a computer program for mixed-effects ordinal regression analysis. *Computer Methods and Programs in Biomedicine, 49*, 229-252.

Heywood, H. B. (1931). On finite sequences of real numbers. *Proc. Roy. Soc. Lon. 4*: 245-253.

Howe, W. G. (1955). "Some contributions to factor analysis." Report No. ORNL-1919. Oak Ridge, TN: Oak Ridge National Laboratory.

Huttenlocher, J. E., Haight, W., Bryk, A. S., & Selzer, M. (1991). Early Vocabulary Growth: Relation to Language Input and Gender. *Developmental Psychology, 27(2)*, 236-249.

Jennrich, R. J., & Schluchter, M. D. (1986). Unbalanced Repeated-Measures Models with Structured Covariance Matrices. *Biometrics, 42*, 805-820.

Jöreskog, K. G. (1979). "Author's adendum to: a general approach to confirmatory factor analysis," in K. G. Jöreskog and D. Sörbom (eds.) Advances in Factor Analysis and Structural Equation Models. Cambridge, MA: Abt Books.

Jöreskog, K. G. and Sörbom, D. (1979). *Advances in factor analysis and structural equation models*. Cambridge, MA: Abt Books.

Jöreskog, K. G., & Sörbom, D. (1995). *LISREL8 User's Manual*. Chicago: Scientific Software International.

Kamata, A. (1998). *Some Generalizations of the Rasch Model: An Application of the Hierarchical Generalized Linear Model*. Unpublished doctoral dissertation, Michigan State University, East Lansing.

Kalaian H. A., & Raudenbush, S. W. (1996). A Multivariate Mixed Linear Model for Meta-Analysis. *Psychological Methods, Vol. 1. No. 3*, 227-335.

Kline, R. B. (1998). *Principles and Practices of Structural Equation Modeling*, New York: Guilford Press.

Lee, V., & Bryk, A. S. (1989). A Multilevel Model of the Social Distribution of High School Achievement. *Sociology of Education, 62,* 172-192.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: Wiley.

Longford, N. T. (1987). Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Random Effects. *Biometrika, 74(4)*, 817-827.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford: Oxford University Press.

Longford, N. T., & Muthén, B. O. (1992). Factor analysis for clustered observations. *Psychometrika 57*, 581-97.

Lord, F. M., & Novick, M. R. (1968). *Statistical Theories of Mental Test Scores*. Menlo Park, Calif: Addison-Wesley.

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.

Muthén, B. O. (1998). *M-Plus User's*. Muthén & Muthén, Los Angeles.

Muthén, B. O. (1998). Second-generation structural equation modeling with a combination of categorical and continuous latent variables: New opportunities for latent class/latent growth modeling. Paper presented at the conference New Methods for the Analysis of Change, Pennsylvania State Univerisity, October 1998.

Muthén, B. O. & Curran, P. J. (1997). General Longitudinal Modeling of Individual Differences in Experimental Designs: A Latent Variable Framework for Analysis and Power Estimation. *Psychological Methods, Vol. 2, No. 4*, 371-402.

Muthén, B. O., Khoo, S. T. & Gustafsson, J. E. (1998). Multilevel latent variable modeling in multiple population. Under review.

Rasch, G. (1960). *Probabilistic Models for some intelligence and attainment tests.* Copenhagen: Danish Institute of Educational Research.

Raudenbush, S. W. (1994). *Fisher Scoring/IGLS and EM algorithm for a variety of two-level models*, College of Education, Michigan State university, Unpublished Manuscript.

Raudenbush, S. W., & Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. *Sociology of Education, 59,* 1-17.

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling Multivariate Effect Sizes. *Psychological Bulletin, Vol. 103, No. 1,* 111-120.

Raudenbush, S. W., Rowan, B., & Kang, S. J. (1991). A Multilevel, Multivariate Model for Studying School Climate With Estimation Via the EM Algorithm and Application to U.S. High-School Data. *Journal of Educational Statistics, Vol. 16, No. 4,* 295-330.

Raudenbush, S. W., & Sampson, R. (1997). *Assessing Direct and Indirect Associations in Multilevel Designs with Latent Variables*. (in press).

Raudenbush, S. W., & Kasim, R. M. (1998). Cognitive Skill and Economic Inequality: Findings from the National Adult Literacy Survey. *Harvard Educational Review, Vol. 68, No. 1, Spring* 33-79.

Raudenbush, S. W., & Sampson, R. (1999). "Ecometrics" Toward a Scientific Ecological Settings, with Application to the Systematic Social Observation. *Sociological Methodology, Vol. 29*, 1-41.

Reckase, M. D. (1991). The Discriminating Power of Items That Measure More Than One Dimension. *Applied Psychological Measurement, Vol. 15, No. 4, December*, 361-373.

SAS Institute (1990). *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume* 1. SAS Institute Inc., Cary, NC, USA.

SAS Institute (1996). *SAS/STAT Software: Changes and Enhancements through Release 6.11*, Cary, NC, USA.

Thum, Y. M. (1997). Hierarchical Linear Models for Multivariate Outcomes. *Journal of Educational and Behavioral Statistics, Spring 1997, Vol. 22, No. 1*, 77-108.

Thurston, L. L. (1947). *Multiple Factor Analysis*. Chicago, IL: University of Chicago Press.

Wald, A. (1950). A note on the identification of economic relations. In T. C. Koopmans, ed., *Statistical Inference in Dynamic Economic Models*. New York: Wiley, pp. 238-244.

Wolfran, S. (1991). *Mathematica: A system for Doing Mathematics by Computer*. Second Edition. Reading, MA: Addison-Wesley.

Willet, J. B., & Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of change. *Pychological Bulletin, 116*, 363-381.

Willet, J. B., & Sayer, A. G. (1996). Growth modeling and covariance structure analysis. In *Advanced Structural Equation Modeling, Issues and Techniques* (Ed. by Marcoulides, G. A. & Schumacker, R. E.) 125-157.