This is to certify that the

thesis entitled

INVESTIGATION OF METHODS OF ANALYZING
HIERARCHICAL DATA

presented by

Boonreang Kajornsin

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Counseling

and Educational Psychology

(Statistics and Research Design)

William B Schmidt

Major professor

Date October 9, 1980

INVESTIGATION OF METHODS OF ANALYZING HIERARCHICAL DATA

By

Boonreang Kajornsin

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling and Educational Psychology

1980

ABSTRACT

INVESTIGATION OF METHODS OF ANALYZING

HIERARCHICAL DATA

By

Boonreang Kajornsin

In recent years researchers have become more cognizant of the problems of analyzing hierarchical data. It has become increasingly evident that efforts to investigate the relationship among educational variables have suffered from a failure to understand complications caused by hierarchical data. When faced with the analysis of hierarchical data many researchers have proposed alternative ways of analyzing such data.

The general purpose of this dissertation was to investigate various alternatives used to analyze hierarchical data by applying them to a set of simulated data. This study extends the regression model presented by Burstein, Linn and Capell (1978) to its multivariate form. The model used to simulate the data is the random effects model. The main assumption used in this model is that there is homogeneity of the within-group regression coefficients. The main concern of this dissertation is to determine which approach gives the best estimates of the between and within regression coefficients in terms of accuracy (least amount of bias) and in terms of precision for various situations. The bias ratio of each estimator was also computed to facilitate comparisons.

Two situations were investigated in this dissertation. The first

situation was one in which there were both individual level predictors which were aggregated to the group level and predictors which were defined only at the group level. The second situation was one in which there were only individual level predictors which could be aggregated. For each situation, three different data sets were generated; first, there were no group level effects; second, group level effects were equal to the individual level effects; third, group level effects were not equal to the individual level effects.

The simulation results showed that all analysis approaches gave the same estimates of the pooled within-group regression coefficients for all six cases with good precision and small bias ratios.

In the situation where there were both individual level predictors which were aggregated to the group level, the group level analysis approach, full model analysis approach and substraction analysis approach all gave essentially the same estimates of the regression coefficients defined for the group level variables. In the case where there was no group level effects, the two stage analysis approach gave better estimates of the regression coefficients defined for the group level variables than for the other three approaches. In the case where the between-group regression coefficients were equal to the pooled within-group coefficients, all four approaches gave essentially the same estimates of the regression coefficients defined for the group level variables. In the case where the between-group regression coefficients were not equal to the pooled within-group regression coefficients and when the intraclass correlations were low (about 0.30) all four approaches gave the same estimates, but when the intraclass correlations were high (about 0.90) the two stage analysis approach did

not give estimates of the regression coefficients as good as those given by the other three approaches.

In the situation where there were only individual level predictors which could be aggregated to the group level, the simulation results showed that for all three cases, the full model analysis approach and the subtraction analysis approach gave exactly the same estimates of the between-group regression coefficients but they were not close to the true parameter values. The group level analysis and Bock application approaches gave estimates of the between-group regression coefficients that were not that different from each other and were also close to the parameters. When the intraclass correlations were high (about 0.90), the group level analysis approach seemed to give better estimates of the between-group regression coefficients, but the Bock application analysis approach gave better estimates when the intraclass correlations were low (about 0.30) in the case where the between-group regression coefficients were not equal to the within regression coefficients. When the between-group regression coefficients were equal to zero, the Bock application analysis approach gave better estimates of the between-group regression coefficients than the group level analysis approach. However, when the between-group regression coefficients were equal to the pooled within group regression coefficients, the two approaches gave essentially the same estimates for the between-group level analysis approach.

ACKNOWLEDGEMENTS

I wish to express my sincere appreciation to my advisor and committee chairman, Professor William H. Schmidt for his assistance, suggestions, and encouragement throughout all phases of my study. Special thanks goes to Dr. Richard Houang, Dr. Robert Floden and Dr. Dennis Gilliland for their help and valuable comments.

Working in the Office of Research Consultation provided me with a most valuable experience which I will never forget. Many thanks to Professor Joe L. Byers, the Director of the Office of Research Consultation, who gave me this job, and to Professor William H. Schmidt, who gave me four years of teaching assistantship experience.

I acknowledge with appreciation the support of the Thai Government which allowed me to pursue my doctoral studies. Special thanks are also extended to Apinya Assavanig for typing part of the rough draft of this document, and to Donna Schmidt, who was kind enough to spend many hours in correcting my English.

Most of all, I wish to thank my husband, Dr. Samnao Kajornsin, for his encouragement, support, understanding and sympathy. Finally, I wish to extend my gratitude to my parents, my brothers, and to Rungson Kajornsin, my son.

TABLE OF CONTENTS

APPENDICES Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

STATEMENT OF THE PROBLEM

In recent years, the problems of analyzing hierarchical data have been well known among researchers. It has become increasingly evident that efforts to investigate the relationship between variables have suffered from a failure to understand complications caused by hierarchical data. Most educational data are hierarchically arranged, i.e., students are grouped into classrooms which are grouped within grade levels and within schools. The schools are also grouped within school districts and these in turn are also grouped within state educational administrations.

Consider the problem of modeling the effects of school structure on student achievement. Suppose we are interested in the effects of some characteristic of school structure on achievement. There is a systematic sorting of families into school districts that produces a correlation of individual student attributes with school characteristics. Therefore, an adequate description of the achievement process must contain both student characteristics and school characteristics.

The practical problem is that the researcher may either analyze individual level data (e.g., regressing individual achievement on student characteristics and school characteristics) or he may analyze school level averages (e.g., regressing average achievement on average individual characteristics and school characteristics). Hannan and Young (1976) have shown that in most realistic situations (i.e., when models are not perfectly specified) results of the two different analyses will be quite dissimilar.

Bidwell and Kasarda (1975) argue that when the question is posed at the school level, the school level regression is most appropriate. Hannan, Freeman and Meyer (1976) point out that researchers seldom adequately specify school-level processes that are anything more than the sum of individual-level processes. The causal arguments are concerned with the impacts on individual students which are composed of school-level outcomes. Consequently, the choice of level is open to question.

Wiley (1973) points out the problem of analyzing data when using large numbers of correlated explanatory variables. He indicates that when variables defined at the level of the individual pupil are aggregated to the level of the school their correlations tend to increase. As a consequence, in the presence of large numbers of such variables, effective analyses are hindered by excessive collinearity (high relations among independent variables). When the number of such collinear variables becomes very large, the effects of individual variables become very difficult to detect. Whenever variables are defined at the school level, the appropriate unit of analysis is the school and the number of degrees of freedom available is limited to the number of schools.

Research on the differences between multiple regression models applied at different levels of aggregated data indicates three things: 1) there are substantial differences in the magnitude of regression coefficients across aggregated levels for specific models; 2) different variables enter the models at different levels; and 3) aggregation of individual characteristics generally inflates the estimated effects of pupil background and thus decreases the likelihood of identifying

teacher and classroom characteristics that are effective. The results cited above are not very comforting for the researcher who wishes to draw conclusions about educational processes at one level but is constrained to analysis at a different level.

When faced with the analysis of hierarchical data many researchers have tried to propose alternative ways of analyzing such data (e.g., Keesling and Wiley, 1974; Cronbach and Webb, 1975; Keesling, 1976; and Burnstein, 1976).

Keesling and Wiley (1974) propose a two stage analysis of hierarchical data. They set out to define a model for disentangling the effects of variables defined solely at the school level from those defined at the level of the pupil.

Cronbach (1975) claims that the overall between-student coefficient from the regression of individual outcome on individual outcome on individual explanatory variables is a composite of the between groups regression coefficient and the pooled within-group regression coefficient. He recommends that between group effects and individual within group effects should be examined separately.

According to Keesling (1976), to obtain the correct estimates of the between school regression coefficient at least two models need to be examined. These two models are a school level model and an individual within school level model. Keesling recommends subtracting the within school regression coefficient from the between school regression coefficient to obtain the correct regression coefficient appropriate to school level effects.

Burstein (1976) proposes an alternative approach by suggesting the examination of determinants of heterogeneity of the within class

slope. He suggests that the first step is to find the specific within class adjusted intercept and slope. The second step is to fit a model at the class level with the adjusted intercept and slope used as outcome variables and the class level explanatory variables used as independent variables.

The general purpose of this dissertation is to investigate various alternatives used to analyze hierarchical data by applying them to a set of simulated data. This study extends the regression model presented by Burstein, Linn and Capell (1978) to its multivariate form. The model used to simulate the data is the random effects model. The main assumption used in this model is the homogeneity of the within group regression, that is, in contrast with Burstein's approach which suggests allowing for the heterogeneity of the within group regressions.

The main concern of this dissertation is to determine which approach gives the best estimates of the between and within regression coefficients in terms of accuracy, least amount of bias and in terms of precision for various situations. In other words, this dissertation is concerned with determining how correctly the alternative procedures tend to work, i.e., how similar the estimated coefficients at the group level are to the known parameter values, and if the conclusions arrived at under each analysis approach are the same.

Two situations are investigated. The first is where there are both individual level predictors which can be aggregated, and predictors defined only at the group level. For example, the predictors could be length of the school day (group level), and average home background (individual level aggregated to the group level). The second situation is where there are only individual level predictors which are aggregated.

For example, the predictors are average home background, and average pretest scores (both individual level variables aggragated to the group level). For each situation, three different populations are investigated; first, there are no group level effects; second, group level effects are equal to the individual level effects; third, group level effects are not equal to the individual level effects.

For the first situation, four analytical approaches will be investigated: a two stage least squares-analysis recommended by Keesling and Wiley, a group level analysis approach using only averages recommended by Cronbach and Webb, a full model analysis approach recommended by Keesling. For the second situation, four approaches will be investigated: the group level analysis approach, an approach based on Bock (1968) using his method of estimating heritable variation in twin studies, the full model analysis approach and the subtraction analysis approach.

The method of investigating these various approaches will involve the use of simulated data which are generated by computer algorithms where the population parameters are known. For each population, fifty samples were generated. By analyzing fifty samples, one can compare 1) the empirical distribution of the estimator for each analysis approach to the others and 2) the empirical standard errors of the parameter estimates. These results can be used to help determine the appropriateness of each of the analyses for different data situations.

CHAPTER II

REVIEW OF THE LITERATURE

Traditionally, in a situation involving heirarchical data, a variety of competing points of view have been cited as justification for the choice of either pupils or groups (classroom, schools, etc.) as the unit of analysis. Hannan (1976) has shown that in most inexact cases (i.e., when the models are not perfectly specified) results of individual level analyses and group level analyses will be quite dissimilar. This finding makes the choice between models extremely important. This section will review the methodologies that some investigators have used to analyze multi-level data.

Cronbach and Webb (1975) reanalyzed a study by G. L. Anderson. Anderson reported finding an interaction of drill and meaningful methods of arithmetic instruction with student ability and achievement. Drill was found to be superior for "overachievers" and meaningful instruction for "underachievers" in 18 fourth-grade classrooms. Pretest measures used in the study were the Minnesota School Ability Test and the Compass Survey Test. Cronbach and Webb argued the importance of separating the regression effects into the between-class and within-class categories. In their reanalysis, separating between-class and within-class regression components of the outcome on aptitude, the Aptitude by Treatment interaction finding disappeared. An apparent interaction in the between-class analysis was dismissed as unreliable. No interactions were found within classes. Finally, the concluded that studies of interactions usually have not been powerful enough to evaluate outcome

on aptitude regressions accurately. Using the class as the unit of analysis, even the rather large Anderson study could not set narrow confidence limits on the regression slopes. They urged investigators collecting data on intact classes to examine between group and within group regressions separately.

Keesling and Wiley (1974) discussed the problem of disentangling the effects of variables defined solely at the level of the school (e.g., length of the school day or the highest degree held by the principal) from those defined at the level of the individual pupil (e.g., home background characteristics). They summarized the model implicit in this situation by:

$$Y_{ij} = \gamma_o + \underline{\gamma}'\underline{Z}_i + \Theta_i + \underline{\beta}'\underline{X}_{ij} + \varepsilon_{ij}$$

where Y_{ij} is the outcome of the $\underline{j}$th pupil in the $\underline{i}$th school, γ_o is an additive constant, $\underline{\gamma}'$ is the vector of adjusted effects of school characteristics on Y, $\underline{Z}_i$ is the vector of school variables for the $\underline{i}$th school, Θ_i is an error component defined at the school level, $\underline{\beta}'$ is the vector of adjusted effects of individual characteristics on Y, $\underline{X}_{ij}$ is the vector of the characteristic of the $\underline{j}$th individual in the $\underline{i}$th school, and ε_{ij} is an error component defined at the individual level.

In the context of hierarchically defined educational data, they proposed three alternatives to obtain appropriate adjusted estimates of the effects at the individual and school levels. The first alternative was to assume that the model was completely specified at the school level, i.e., all of the school variables relevant to the outcome are included in the model. Then the covariance $(\Theta_i, \overline{\underline{X}}_i)$ is equal to zero, where $\overline{\underline{X}}_i$ is the mean of $\underline{X}_{ij}$ for the $\underline{i}$th school. This model

implies that individual level variables have direct impact on outcomes only at the level of the individual; their effects at the school level are mediated through other variables defined at the level of the school.

The second alternative was that if all the mediating variables at the school level were not specified in the model, then the covariance $(\Theta^*_i, \underline{\bar{X}}_i)$ was not equal to zero where Θ^*_i was the residual from the measured school variables. In this case, the fitting of the model will produce a biased estimate of $\underline{\beta}$. This source of bias may, however, be eliminated by performing an analysis based on the variation within schools. This may be done by subtracting the relevant school means (school effect values) for the criterion variable and for each of the pupil level explanatory variables from each of the individual values for these variables. An analysis performed using these deviated values will be adjusted for all sources of variation among schools. The covariance matrix of the deviated values is called the pooled within school covariance matrix. If this covariance matrix is computed for all individually defined variables and used as the basis for the regression of the outcome on the $\underline{X}_{ij}$, the resulting estimate of $\underline{\beta}$ will not be biased by specification errors at the school level.

After the adjusted effects of the individual level variables are found, the average effect value for each school, aggregated over all the individual pupils, may be subtracted from the criterion mean for each school. Analyses using the school as a unit with variables defined at the level of school as the independent variables and the modified criterion means as the dependent variable will produce estimates of the effect of the school variables adjusted for the effects of individually defined variables. The model at the school level

becomes:

$$\bar{Y}_i - \hat{\underline{\beta}}' \bar{\underline{X}}_i = \gamma_o + \underline{\gamma}' \underline{Z}_i + \phi_i$$

where $\bar{Y}_i$ is an achievement mean for i<u>th</u> school, $\hat{\underline{\beta}}' \underline{X}_i$ is the estimated average effect value for i<u>th</u> school, γ_o is the constant, $\underline{\gamma}'$ is the vector of the adjusted effects for the school variables, and ϕ_i is an error component at the school level. Using this model, the analysis will produce unbiased estimates of γ_o and $\underline{\gamma}$ in the absence of specification error.

The third alternative was that if there was some specification bias at the level of the school-defined variables (i.e., some important variables are missing) then the covariance (Θ_i, $\bar{\underline{X}}_i$) is not equal to zero and the covairance (ϕ_i, $\bar{\underline{X}}_i$) is not equal to zero, either. Some of the biases in the estimate of $\underline{\gamma}$ can be removed by including the sum of the average effect values ($\hat{\underline{\beta}}' \bar{\underline{X}}_i$) of the individually defined varialbes as another variable in the school level analysis. The model then becomes as follows:

$$\bar{Y}_i = \gamma_o + \underline{\gamma}' \underline{Z}_i + \lambda \, (\hat{\underline{\beta}} \bar{\underline{X}}_i) + \phi_i$$

This technique allows the partial removal of some of the additional bias due to the omission of relevant school level variables to the extent that the sum of these average effect values is correlated with the omitted variables.

Rock, Baird, and Linn (1972) studied the interaction between college effects and students' aptitudes. They claimed that their approach was designed to find groups of colleges that are about equally effective for students with various levels of initial performance. Then the characteristics of the identified criterion groups were compared to see which characteristics were related to the relative

effectiveness of the groups. Their method attempted to provide an intuitively simple approach which identified both overall college effects and effects which interact with student ability. Specifically, four steps were carried out: 1) all within school regression lines were computed, i.e., Graduate Record Examination (GRE), area tests were regressed on the College Entrance Examination Board, Scholastic Aptitude Test (SAT) scores within school; 2) Ward's (1963) hierarchical clustering technique was applied to group schools in the basis of the similarity of their regression lines; 3) multiple group discriminant functions using the estimates of the regression parameters as the group discriminants were computed to test whether the newly formed groups differed with respect to their pooled regression lines; and 4) discriminant functions using college descriptive variables as the group discriminants were then computed. This method thus identified criterion clusters of colleges that differed in effectiveness by clustering on the slope, the mean SAT scores of the students, and the intercept. Therefore, one can identify and group colleges that have different levels of initial ability. Then the simultaneous evaluation of the college along with the relative slopes of their pooled within group regression lines indicated the college characteristics which are associated with overall as well as differential effectiveness.

Burstein (1976) discussed two examples of multi-level analyses found in studies by Rock, Baird and Linn (1972) about the interaction of student aptitude and college characteristics and by Keesling and Wiley (1974) in which they reanalyzed a subset of the Coleman data. He stated that each method has certain merits and certain drawbacks. The Keesling and Wiley approach provided effect parameters more nearly

mirroring the structural form of school effects than the Rock, Baird and Linn approach or the usual single level analysis models. Burstein's concern was that the Keesling and Wiley approach fails to adequately reflect the effects of between class differences in slopes. Moreover, treating the resulting clusters as groups in a discriminant analysis as Rock, Baird, and Linn did discarded any metric differences existing among the clusters and thereby eliminated the possibility of describing school effects in structural terms. The use of discriminant groups results in some loss in generalizability of findings that should be avoided. In the same paper Burstein also criticized Cronbach's approach that recommended analyzing between class and within class separately when intact classrooms are sampled. Burstein said that the between class and within class analyses did not remove the need for concern about homogeneity of regression.

Burstein proposed an alternative multilevel analysis atragety that consisted of two stages, as follows:

1. perform within class regression (not pooled) of outcomes on input, and
2. use the parameters (α,β) from the within class regressions as "outcomes" in a between class analysis.

Burstein claimed that his strategy combined certain features of approaches by Keesling and Wiley and by Rock, Baird and Linn. The technique of using the within class parameter estimates as outcomes should lead to more sensitive interpretation of effects and clearer policy implications of the findings.

Burstein and Miller (1979) stated that because of its hierarchical organization, the effects of schooling on individual pupil performance

can exist both between and within the levels of the educational system. Moreover, analyses at different levels address different questions and thus analyses conducted at a single level were inherently inadequate. While analyses of the relationships between "treatment" dimensions and the mean outcomes of groups often provide useful information, important differences in within group processes may be obscured. These within group processes may arise due to group composition (e.g., ability level and mixture affecting participation patterns), differential allocation of instructional resources among the members of the group (e.g., the grouping and pacing features of reading instruction), or differential reactions of group members to the same instructional treatment (aptitude-treatment interactions).

If important group-to-group differences in within group processes exist then the use of group means as the only indicator of group outcomes will result in misleading estimates of group (teacher, class, treatment) effects.

Burstein and Miller's interest in alternative measures of group outcomes has concentrated on the properties of the within-group slopes from the regression of outcome on input. They have argued that within group slopes are group level indicators of within group processes. Their reason for considering slopes as outcomes was that there may be instructional effects on the within group regression of outcomes on input, whether there were instructional effects were present, the analysis should attempt of isolate instructional process and practice variables that were associated with slope variation. If such variables can be found and alternative explanations cannot be ruled out then variation in slopes becomes an important source of information for

researchers and policy makers, especially when considered along with effects on other group level outcomes.

Keesling (1976) presented a model for analysis at two levels of aggregation (e.g., pupil and school). The multivariate random effects model for this situation is:

$$\underline{Y}_{ij} = \underline{\mu} + \underline{a}_i + \underline{e}_{ij} \qquad \begin{matrix} i = 1,2,.....,k \\ j = 1,2,.....,n \end{matrix}$$

all vectors are p x 1. This implies $\Sigma_y = \Sigma_a + \Sigma$ assuming that there are k groups of n units, each unit having measures on p variables.

The above model, adopted from Schmidt (1969), was comprehensive in that it permitted the estimation of effects and their standard errors at both levels of aggregation simultaneously. Keesling, however, did not analyze the data by using Schmidt's procedure.

Two sets of data were presented. One set dealt with data constructed to a particular specification. The second set dealt with real data of a two-level nature. He analyzed data under three models. The first model used pupil post-test score as the dependent variable, ignoring the group structure in the data, and pretest, SES, average pretest, average SES and hours per month of principal absence as predictors. The second model used school mean post-test as the dependent variable, average pretest, average SES and hours per month of principal absence as predictors. The third model used pupil posttest score within school as the dependent variable with pretest and SES as predictors. The results suggested that in order to obtain both the correct parameter estimates and the correct standard errors, it is necessary to perform at least two analyses. The first model gave the correct parameter estimates, but it did not partition the residual sum

of squares by level of effect. The second model gave the aggregate level standard errors, but the parameter estimates were the sum of the between and within effects. The third model obtained the appropriate estimates and standard errors for the within school effects. The second and third model may be combined to produce correct estimates of the between school effects by subtracting the within school estimates from the between school estimates.

CHAPTER III

ALTERNATIVE APPROACHES FOR ANALYZING HIERARCHICAL DATA

Of the different alternatives proposed to analyze the hierarchical data, four were selected for comparison in the present study. The two stage analysis approach which was recommended by Keesling and Wiley, group level analysis approach which was recommeded by Cronbach and Webb, full model and subtraction analysis approaches which were recommended by Keesling, and Bock application analysis approach will be discussed in this chapter.

Consider the following general situation where person j is a member of group i. The person has a set of scores, $\underline{X}_{ij}$ and Y_{ij}. Also available are a set of explanatory variables defined only at the group level which is denoted as $\underline{Z}_i$. The relationship of $\underline{X}_{ij}$ and $\underline{Z}_i$ to Y_i can be decomposed into between group and within group components as given in equation (3-1).

$$(3\text{-}1) \qquad Y_{ij} = \mu_y + \underline{\beta}_a' (\underline{\mu}_{x_i} - \underline{\mu}_x) + \underline{\beta}_z' (\underline{Z}_i - \underline{\mu}_z) + \delta_i + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + (\underline{\beta}_i - \underline{\beta})'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + \varepsilon_{ij}$$

where μ_y, $\underline{\mu}_z$ and $\underline{\mu}_x$ represent the population means, $\underline{\mu}_{x_i}$ represents the ith group population mean, $\underline{\beta}_a$ denotes the between-group regression coefficients for the individual level variables, $\underline{\beta}_z$ represents the regression coefficients defined for the group level variables, $\underline{\beta}$ represents the pooled within-group regression coefficients for the

individual level variables, and $\underline{\beta}_i$ represents the specific within-group regression coefficients for group i for the individual level variables. The δ_i and ϵ_{ij} represent the error at the group level and at the individual level, respectively.

This study will deal with the case where all within group slopes are equal, resulting in $(\underline{\beta}_i - \underline{\beta})$ being equal to zero. The model for the first simulated case is:

$$(3\text{-}2) \quad Y_{ij} = \mu_y + \underline{\beta}_a' (\underline{\mu}_{x_i} - \underline{\mu}_x) + \underline{\beta}_z' (\underline{Z}_i - \underline{\mu}_z) + \delta_i + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + \epsilon_{ij}$$

Let $\underline{a}_{x_i} = \underline{\mu}_{x_i} - \underline{\mu}_x$

$\underline{a}_{z_i} = \underline{Z}_i - \underline{\mu}_z$

$\underline{a}_{x_{ij}} = \underline{X}_{ij} - \underline{\mu}_{x_i}$

The equation (3-2) can be rewritten as equation (3-3):

$$(3\text{-}3) \quad Y_{ij} = \mu_y + \underline{\beta}'\underline{a}_{x_i} + \underline{\beta}'\underline{a}_{z_i} + \delta_i + \underline{\beta}'\underline{a}_{x_{ij}} + \epsilon_{ij}$$

This implies that the variance of Y is:

$$(3\text{-}4) \quad \text{Var}(Y) = \underline{\beta}_a'\Sigma_a^x\underline{\beta}_a + \underline{\beta}_z'\Sigma_a^z\underline{\beta}_z + \sigma_a^2 + \underline{\beta}'\Sigma^x\underline{\beta} + \sigma^2$$

where Σ_a^x is the between level variance-covariance matrix of $\underline{X}$, Σ_a^z is the between level variance-covariance matrix of $\underline{Z}$, Σ^x is the within level covariance matrix of $\underline{X}$, σ_a^2 is the error variance defined at the group level and $\sigma^2$is the error variance defined at the individual level.

Then there are only individual level explanatory variables, the model is:

$$(3\text{-}5) \quad Y_{ij} = \mu_y + \underline{\beta}_a' (\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i + \underline{\beta}' (\underline{X}_{ij} - \underline{\mu}_{x_i}) + \epsilon_{ij}$$

And the variance of Y is:

$$(3\text{-}6)\quad \mathrm{Var}(Y) = \underline{\beta}_a'\Sigma^x_a\underline{\beta}_a + \sigma^2_a + \underline{\beta}'\Sigma^x\underline{\beta} + \sigma^2.$$

The five alternative analysis approaches (two-stage analysis, group level analysis, full model analysis, subtraction analysis, and Bock application analysis) that are investigated in this dissertation can be related to the models as given in equations (3-2) and (3-5) for the first and second situation, respectively. In the following pages, the procedures of each alternative approach is discussed.

Two Stage Analysis Approach

The two stage analysis approach was recommended by Keesling and Wiley (1974). Wiley mentions that one of the problems in the analysis of multi-level data has been separation of the effects of the aggregated variables into parts reflecting their individual level effects on one hand, and their effects via school climate and organization on the other. One way to describe an appropriate method of analysis of hierarchical data is in terms of the general notions of statistical confounding and control. If we wish to assess the impact of how one explanatory variable is correlated with another one, then if we ignore the second, we will attribute to the first not only its effect, but also a spurious effect which is due to the correlation between it and the second, and the effect of the second. If we utilize an appropriate method of analysis which takes into account the second variable, i.e., its effects and its relationship to the first, we may obtain an adjusted assessment of the effect of the first variable which is not confounded by the second.

Keesling and Wiley set out to define a model for disentangling the effects of variables defined solely at the school level from those

defined at the level of the pupil. The process of disentanglement involves two stages. The first stage adjusted the effects of individual background characteristic on outcome for the effects of the schools in which the individuals receive instruction. The second stage used the adjusted effects of individual level variables aggregated over pupils within schools to determine the adjusted effects of school level variables. In practice, they carried out the following:

1. Determine the pooled within-school slopes under equation (3-7).

$$(3\text{-}7) \quad Y_{ij} = \mu_{y_i} + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + \varepsilon_{ij}$$

$i = 1, 2, \ldots, k; \quad j = 1, 2, \ldots, n$

where Y_{ij} is the outcome of the jth subject in the ith school, μ_{y_i} is the population mean of the ith school, $\underline{X}_{ij}$ is the vector of explanatory variables of the jth subject in the ith school, and $\underline{\beta}$ is the vector of pooled within-school slopes.

An analysis using the school mean deviated values of both explanatory and criterion variable will effectively "control" or adjust for all sources of variation among schools. The covariance matrix of the deviated values is called the pooled-within school covariance matrix. If this covariance matrix is computed for all individually defined variables and used as the basis for the regression of the outcome on the set of explanatory variables, the resulting estimates of $\underline{\beta}$ will not be biased by specification errors at the school level.

2. Find the mean predicted outcome for each school.

$$(3\text{-}8) \quad \hat{\mu}_{y_i} = \hat{\mu}_y + \hat{\underline{\beta}}'(\hat{\underline{\mu}}_{x_i} - \hat{\underline{\mu}}_x)$$

where $\hat{\mu}_{y_i}$ is the mean predicted outcome for the ith school.

3. Fit a model at the school level regressing the observed school

mean outcome on school level explanatory variables and predicted school mean outcomes,

$$(3\text{-}9) \quad \mu_{y_i} = \mu_y + \underline{\beta}_z'(\underline{Z}_i - \underline{\mu}_z) + \lambda\mu_{y_i} + \delta_i$$

where $\underline{\beta}_z$ is the vector of adjusted effects of the school level variables, $\underline{Z}_i$ is the vector of the school level variables, δ_i is the error defined at the group level, λ is the coefficient allowing for partial removal of some of the additional bias due to the omission of relevant school level variables (to the extent that the sum of these average effect values is correlated with the sum of the average individual level effect values represented in μ_{y_i}). If all relevant school level variables are included, then λ will be equal to one.

Group Level Analysis Approach

Cronbach (1976) mentions that in the situation where pupil j is a member of group i, β_t, the overall between student coefficient from the regression of Y_{ij} on X_{ij},

$$(3\text{-}10) \quad Y_{ij} = \mu_y + \beta_t(X_{ij} - \mu_x) + \varepsilon_{ij}$$

has been shown by Duncan, Cuzzort, and Duncan (1961) to be a composite of β_a, the between group regression coefficient and β, the pooled within-group coefficient;

$$(3\text{-}11) \quad \beta_t = \eta_x^2\beta_a + (1 - \eta_x^2)\beta$$

where η_x^2 is the correlation ratio of X.

$$(3\text{-}12) \quad \eta_x^2 = 1 - \frac{\Sigma\Sigma_{ij}(X_{ij} - \mu_{xi})^2}{\Sigma\Sigma_{ij}(X_{ij} - \mu_x)^2} \quad .$$

Cronbach indicated that analyses at the group level and the individual level give conflicting descriptive results because they speak to different substantive questions. The investigator who wants to know the relationship between two variables is not asking a clear question until he tells whether the group or individual level relationship is the one of interest. He recommended that between group effects and individual within group effects should be examined separately. He proposed the following:

1. Between groups:

$$(3\text{-}13) \quad \mu_{y_i} = \mu_y + \underline{\beta}_z'(\underline{Z}_i - \underline{\mu}_z) + \underline{\beta}_a'(\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i$$

where the $\underline{\beta}_z$ is the effect of school level variables on mean outcomes, and $\underline{\beta}_a$ is the between groups effect that reflects any consistent tendency of higher-X groups to do better or worse than others on the outcome measure.

2. Pooled within groups:

$$(3\text{-}13) \quad Y_{ij} = \mu_{y_i} + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + \varepsilon_{ij}$$

where $\underline{\beta}$ is the common within-group effect that reflects the tendency for students above the group average to outperform or underperform the rest of the group.

Subtraction Analysis Approach

In the situation where subject j is nested within group i, Keesling (1977) analyzed constructed data to show how well ordinary least square estimators can retrieve the information. He analyzed the data under two models, as follow:

1. The group level model uses group mean outcome as the independent variable:

$$(3\text{-}15) \quad \mu_{y_i} = \mu_y + \underline{\beta}'_z(\underline{Z}_i - \underline{\mu}_z) + \underline{\beta}^{*\prime}_a(\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i$$

Keesling claimed that this model gives the aggregated level standard errors, but the parameter estimates are the sum to the between and within effects.

2. The within group model is the model that uses the individual level outcome variable within groups as the dependent variable. According to Keesling, this model obtains the appropriate estimates and standard errors for the within group effects.

$$(3\text{-}16) \quad Y_{ij} = \mu_{y_i} + \underline{\beta}' (\underline{X}_{ij} - \underline{\mu}_{x_i}) + \varepsilon_{ij}$$

Keesling concluded that to obtain the correct estimates of the between school effects at least these two models need to be performed and then substract the within group estimates from the between group estimates. That is,

$$(3\text{-}17) \quad \underline{\beta}_a = \underline{\beta}^*_a - \underline{\beta}$$

where $\underline{\beta}_a$ is the correct between group effects, $\underline{\beta}^*_a$ is the estimate of the between groups effects using the group level data and $\underline{\beta}$ is the within group effect.

The Full Model Analysis Approach

The full model is the model that uses the individual level outcome variable as the dependent variable. The explanatory variables are: 1) the variables defined at the individual level but which can also be aggregated, 2) the means of the variables defined at the individual level, and 3) the variables defined at the group level only. The model is shown in equation (3-18).

$$(3\text{-}18)\quad Y_{ij} = \mu_y + \underline{\beta}_z' (\underline{Z}_i - \underline{\mu}_z) + \underline{\beta}_a' (\underline{\mu}_{x_i} - \underline{\mu}_x) + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_x) + \varepsilon_{ij}$$

where μ_y, $\underline{\mu}_z$ and $\underline{\mu}_x$ are the population means, $\underline{\mu}_{x_i}$ is the ith group population mean, $\underline{\beta}_a$ represents the between-group regression coefficients for the aggregated individual level variables, $\underline{\beta}_z$ represents the regression coefficients for the group level variables, $\underline{\beta}$ represents the pooled within-group regression coefficients for the individual level variables, and ε_{ij} represents the error defined at the individual level.

Keesling (1977) at one time analyzed the heirarchical data under the full model. He mentioned that this model gave the correct parameter estimates, but it did not partition the residual sum of squares by the level of effect.

Bock Application Analysis Approach

In the situation where there are students nested within schools and the school is a random variable, the model is the random effects model. In this dissertation, there is one dependent measure and two antecedent measures for each subject; the random effects model is:

$$\underline{W}_{ij} = \underline{\mu} + \underline{a}_i + \underline{e}_{ij} \quad , \quad \begin{matrix} i = 1, 2, \ldots, k. \\ j = 1, 2, \ldots, n. \end{matrix}$$

where all vectors are 3x1 in this application, $\underline{W}_{ij}$ is the response vector representing the dependent, and antecedent measures, $\underline{\mu}$ is vector of the population means on each measure, $\underline{a}_i$ and $\underline{e}_{ij}$ are the random vectors assumed to be multivariate normally and independently distributed with zero mean vectors and covariance matrices Σ_a and Σ respectively. The above model implies $\Sigma_w = \Sigma_a + \Sigma$, where Σ_w is the total variance

covariance matrix, Σ_a is the between school variance matrix, and Σ is the within school variance covariance matrix.

The use of the Bock application approach is to provide an estimate of Σ_a which is at least a positive semi-definite variance covariance matrix, and then from this matrix to estimate the group level regression coefficients. Bock's method is presented in the context of twin studies and is used to estimate the component of heritable variation. A more detailed description of this approach can be found in Bock (1968).

Under the random effects model, the expected value of the mean square matrix between schools is $\Sigma + n\Sigma_a$, and the expected value of the mean square matrix within schools is Σ.

$$\text{Let } S_a = \Sigma + n\Sigma_a$$
$$S = \Sigma$$

Then for a symmetric positive definite matrix S and a symmetric positive definite matrix S_a, it is possible to find a nonsingular transformation T such that

$$T'S_aT = \Phi \tag{3-19}$$
$$T'S\,T = I \tag{3-20}$$

where Φ is diagonal with positive diagonal elements, and I is an identity matrix. The columns of T are the solution of a system of homogeneous equations of the form:

$(S_a - \phi_1 S)t = 0$, $1 = 1, 2, 3$, and ϕ_1 is a root of

$|S_a - \Phi S| = 0$.

In practice, the estimate of S is the mean square matrix within schools, that is obtained from the equation (3-21).

$$S = \frac{1}{k(n-1)} \sum_{i}^{k} \sum_{j}^{n} (\underline{W}_{ij} - \underline{W}_i)(\underline{W}_{ij} - \underline{W}_i)' \tag{3-21}$$

where $\underline{W}_{ij}$ is the individual response vector, $\underline{W}_i$ is the group mean vector, k is the number of groups and n is the number of subjects in each group.

$$\underline{W}_{ij} = \begin{bmatrix} Y_{ij} \\ \underline{X}_{ij} \end{bmatrix}, \quad \underline{W}_i = \begin{bmatrix} \bar{Y}_i \\ \underline{\bar{X}}_i \end{bmatrix}$$

The estimate of S_a is the mean square matrix between schools that is obtained from the equation (3-22).

$$(3\text{-}22) \quad S_a = \frac{1}{k-1} \; n\sum_{i}^{k} (\underline{W}_i - \underline{W})(\underline{W}_i - \underline{W})'$$

where $\underline{W}$ is the grand mean vector. $\underline{W} = \begin{bmatrix} \bar{Y} \\ \bar{X} \end{bmatrix}$.

From equations (3-19) and (3-20),

$$\begin{aligned} T'S_aT - T'ST &= \Phi - I \\ T'(S_a - S)T &= \Phi - I \\ T'n\Sigma_aT &= \Phi - I \\ \Sigma_a &= [(T^{-1})'(\Phi - I)T^{-1}]/n \end{aligned}$$

Practically, for the elements in the columns of T the discriminant function coefficients are substituted and for the elements of Φ the corresponding significant canonical variances Φ_l (l = 1, 2, . . . , s) and p -s unities are substituted (p is the dimension of T). This estimate has the following properties. Because the elements of the diagonal matrix (Φ -I) are non-negative, it can be expressed as the product of a matrix and its transpose and is therefore positive semi-definite. Its rank is s and its nullity is p - s. When all of the canonical variances are significant (s = p),

$$\begin{aligned} \hat{\Sigma}_a &= [(T^{-1})'(T'S_aT - T'ST)T^{-1}]/n \\ &= [S_a - S]/n \end{aligned}$$

where S_a is the estimate of mean square matrix between schools and S is the estimate of mean square matrix within school.

The between school covariance matrix (Σ_a) estimated in this way and guaranteeing positive semi-definiteness can then be used to estimate the between school regression coefficient.

CHAPTER IV

SIMULATION PROCEDURE

Simulation procedures were used in this study to generate the data. The use of simulated data enables us to determine which method of analyzing hierarchical data gives the best estimator, in terms of accuracy and precision, of the parameters under various situations. The bias ratio of each estimator was also computed to facilitate comparisons. Two situations were investigated in this dissertation. The first situation was one in which there were both individual level predictors which can be aggregated to the group-level and predictors defined only at the group-level. The second situation was one in which there were only individual-level predictors which can be aggregated. For each situation, three different data sets will be generated. These are described in the following ways:

1. No group level effect. The between-group regression coefficient is set to zero, but the within-group regression coefficient is non-zero. This case implies that there is no group level effect (i.e., $\underline{\beta}_a = 0$).

2. Group level effect is equal to the individual level effect. The between group regression coefficient is not equal to zero but is equal to the within-group regression coefficient. This case implies that there is a group level effect, and that the group level effect is equal to the within level group effect (i.e., $\underline{\beta}_a = \underline{\beta} \neq 0$).

3. Group level effect is not equal to the individual level effect.

The between group regression coefficient is neither equal to zero, nor equal to the within-group regression coefficient. This case implies that there is a group level effect but it is not equal to the within-group effect.

Description of Population Parameters

The data generated for the present study were from a multivariate normal distribution with a mean vector $\underline{\mu}$ and a covariance matrix $\Sigma + \Sigma_a$, where Σ is the within covariance matrix, and Σ_a is the between covariance matrix. The between-groups and within-groups regression coefficients ($\underline{\beta}$ and $\underline{\beta}_a$), the within and between covariance matrix of individual level predictors ($\Sigma^{(x)}$ and $\Sigma_a^{(x)}$), the between covariance matrix of predictors defined at the group level only ($\Sigma^{(z)}$), the between covariance matrix of predictors defined at the individual level and at the group level ($\Sigma^{(xz)}$), the error variance at the individual level (σ^2), the error variance at the group level (σ_a^2), and the population mean ($\underline{\mu}$) were specified in advance.

The study by Keesling and Wiley (1974) was used as a guide to choose parameter values which would be reasonable. Three population covariance matrices were constructed based on the model (3-2) and (3-5) for the first and second situations, respectively. The six possible total covariance matrices ($\Sigma + \Sigma_a$) were derived from the 3x2 crossed design of possible combinations of population parameters and situations (see Table 4-1). Fifty samples of 1,500 subjects each were generated for each cell in Table 4-1. In each sample there were fifty schools with thirty subjects within each school. The structure of the within covariance matrix (Σ) and the between covariance matrix (Σ_a) for the

Table 4-1

3x2 Design of Populations Defining the Structure of $(\Sigma + \Sigma_a)$

	Both individual level and school level variables.	Individual level variables only.
$\underline{\beta}_a = \underline{0}$, $\underline{\beta} \neq \underline{0}$	I-A	II-A
$\underline{\beta}_a = \underline{\beta} \neq \underline{0}$	I-A	II-B
$\underline{\beta}_a \neq \underline{\beta} \neq 0$	I-C	II-C

first and second situations are shown in Table 4-2. The vector of populations means ($\underline{\mu}$), pooled within-group regression coefficient ($\underline{\beta}$), between-groups regression coefficient ($\underline{\beta}_a$), error variance at the individual level (σ^2), and at the group level (σ_a^2) for both situations are given in Tables 4-3, and 4-4. The numerical values of $\Sigma^{(x)}$, $\Sigma_a^{(x)}$, $\Sigma_a^{(z)}$, $\Sigma_a^{(xz)}$, Σ, and $\Sigma + \Sigma_a$ are given in Tables 4-5, 4-6, and 4-7.

The intraclass correlations of variables Y and $\underline{X}$ are quite high (about 0.6052 for Y and 0.6547 for $\underline{X}$) in population I-C. In population II-C, the intraclass correlations of Y and $\underline{X}$ variables are all about 0.98. Therefore, in order to check whether the analysis approches give the same result in the situation where the intraclass correlations are not as high as the first set of data, a second set of data for populations I-C and II-C were generated with a new set of parameters as shown in Table 4-8 which have intraclass correlations of about 0.30. The numerical values of $\Sigma^{(x)}$, $\Sigma_a^{(x)}$, $\Sigma_a^{(z)}$, $\Sigma_a^{(xz)}$, Σ, Σ_a,

Table 4-2

Population Compositions of Σ and Σ_a

Situation	Within Covariance (Σ)			Between Covariance (Σ_a)		
I	$\underline{\beta}'\Sigma^{(x)}\underline{\beta} + \sigma^2$	(Symmetric)		$\underline{\beta}_a'\Sigma_a^{(x)}\underline{\beta}_a + 2\underline{\beta}_a'\Sigma_a^{(xz)}\underline{\beta}_z + \underline{\beta}_z'\Sigma_a^{(z)}\underline{\beta}_z + \sigma_a^2$	(Symmetric)	
	$\Sigma^{(x)}\underline{\beta}$	$\Sigma^{(x)}$		$\Sigma_a^{(x)}\underline{\beta}_a + \Sigma_a^{(xz)}\underline{\beta}_z$	$\Sigma_a^{(x)}$	
	$\underline{0}$	$\underline{0}$	$\underline{0}$	$\Sigma_a^{(zx)}\underline{\beta}_a + \Sigma_a^{(z)}\underline{\beta}_z$	$\Sigma_a^{(xz)}$	$\Sigma_a^{(z)}$
II	$\underline{\beta}'\Sigma^{(x)}\underline{\beta} + \sigma^2$	(Symmetric)		$\underline{\beta}_a'\Sigma^{(x)}\underline{\beta}_a + \sigma_a^2$	(Symmetric)	
	$\Sigma^{(x)}\underline{\beta}$	$\Sigma^{(x)}$		$\Sigma_a^{(x)}\underline{\beta}_a$	$\Sigma_a^{(x)}$	

Table 4-3

Parameter Values for the First Situation

Case	$\underline{\mu}$	$\underline{\beta}$	$\underline{\beta}_a$	$\underline{\beta}_z$	σ^2	σ_a^2
I-A	$\begin{bmatrix} 12.0810 \\ 1.3491 \\ 2.4587 \\ 6.8660 \\ 1.0511 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	$\begin{bmatrix} 4.08 \\ 2.15 \end{bmatrix}$	0.5276	0.0812
I-B	$\begin{bmatrix} 12.0810 \\ 1.3491 \\ 2.4587 \\ 6.8660 \\ 1.0551 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 4.08 \\ 2.15 \end{bmatrix}$	0.5276	0.8972
I-C	$\begin{bmatrix} 12.0810 \\ 1.3491 \\ 2.4587 \\ 6.8660 \\ 1.0511 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 1.45 \\ 0.89 \end{bmatrix}$	$\begin{bmatrix} 4.08 \\ 2.15 \end{bmatrix}$	0.5276	0.9547

Table 4-4

Parameter Values for the Second Situation

Case	$\underline{\mu}$	$\underline{\beta}$	$\underline{\beta}_a$	σ^2	σ_a^2
II-A	$\begin{bmatrix} 25.3810 \\ 12.5912 \\ 11.4587 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 0 \\ 0 \end{bmatrix}$	0.5276	20.7321
II-B	$\begin{bmatrix} 25.3810 \\ 12.5912 \\ 11.4587 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	0.5276	32.7341
II-C	$\begin{bmatrix} 25.3810 \\ 12.5912 \\ 11.4587 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 1.45 \\ 0.89 \end{bmatrix}$	0.5276	22.3454

Table 4-5

Population Covariance Matrices of the Predictor Variables

Covariance Matrices	Situation I	Situation II
$\sum_a^{(x)}$	$\begin{bmatrix} 0.0912 & 0.1901 \\ 0.1901 & 2.4775 \end{bmatrix}$	$\begin{bmatrix} 0.0912 & 0.1901 \\ 0.1901 & 2.4775 \end{bmatrix}$
$\sum_a^{(x)}$	$\begin{bmatrix} 0.1729 & 0.1400 \\ 0.1400 & 0.3746 \end{bmatrix}$	$\begin{bmatrix} 14.7149 & 2.9871 \\ 2.9871 & 25.8970 \end{bmatrix}$
$\sum_a^{(z)}$	$\begin{bmatrix} 0.0072 & 0.0007 \\ 0.0007 & 0.0009 \end{bmatrix}$	Not Applicable
$\sum_a^{(xz)}$	$\begin{bmatrix} 0.0159 & 0.0065 \\ 0.0260 & 0.0088 \end{bmatrix}$	Not applicable

Table 4-6

Population Covariance Matrices Under the First Situation

Case	Within Covariance (Σ)					Between Covariance (Σ_a)					Total Covariance ($\Sigma+\Sigma_a$)				
	1.763		(Symmetric)			0.218		(Symmetric)			1.890		(Symmetric)		
	0.292	0.091				0.079	0.173				0.370	0.264			
I-A	1.274	0.190	2.478			0.125	0.140	0.375			1.399	0.330	2.852		
	0	0	0	0	0	0.031	0.016	0.026	0.007		0.031	0.016	0.026	0.007	
	0	0	0	0	0	0.005	0.007	0.009	0.001	0.001	0.005	0.007	0.009	0.001	0.001
	1.673		(Symmetric)			2.885		(Symmetric)			4.557		(Symmetric)		
	0.292	0.091				0.561	0.173				0.853	0.264			
I-B	1.274	0.190	2.478			0.599	0.140	0.375			1.873	0.330	2.852		
	0	0	0	0		0.079	0.016	0.026	0.007		0.079	0.016	0.026	0.007	
	0	0	0	0	0	0.024	0.007	0.009	0.001	0.001	0.024	0.007	0.009	0.001	0.001
	1.673		(Symmetric)			2.564		(Symmetric)			4.234		(Symmetric)		
	0.292	0.091				0.454	0.173				0.746	0.264			
I-C	1.274	0.190	2.478			0.661	0.140	0.375			1.935	0.330	2.852		
	0	0	0	0		0.077	0.016	0.026	0.007		0.077	0.016	0.026	0.007	
	0	0	0	0	0	0.022	0.007	0.009	0.001	0.001	0.022	0.007	0.009	0.001	0.001

Table 4-7

Population Covariance Matrices Under the Second Situation

Case	Within Covariance (Σ)	Between Covariance (Σ_a)	Total Covariance ($\Sigma+\Sigma_a$)
II-A	1.673 (Symmetric) 0.292 0.091 1.274 0.190 2.478	20.732 (Symmetric) 0 14.715 0 2.987 25.897	22.405 (Symmetric) 0.292 14.806 1.274 3.177 28.375
II-B	1.673 (Symmetric) 0.292 0.091 1.274 0.195 2.478	134.411 (Symmetric) 38.185 14.715 15.844 2.987 25.897	136.084 (Symmetric) 38.476 14.806 17.118 3.177 28.375
	1.673 (Symmetric) 0.292 0.091 1.274 0.190 2.478	81.506 (Symmetric) 23.995 14.715 27.380 2.987 25.897	83.179 (Symmetric) 24.287 14.806 28.653 28.375

Table 4-8

Parameter Values of the Second Set of Data

Case	$\underline{\mu}$	$\underline{\beta}$	$\underline{\beta}_a$	$\underline{\beta}_z$	σ^2	σ_a^2
I-C	$\begin{bmatrix} 12.0810 \\ 1.3491 \\ 2.4587 \\ 6.8660 \\ 1.0511 \end{bmatrix}$	$\begin{bmatrix} 2.53 \\ 0.32 \end{bmatrix}$	$\begin{bmatrix} 1.45 \\ 0.89 \end{bmatrix}$	$\begin{bmatrix} 4.08 \\ 2.15 \end{bmatrix}$	100.2311	22.3454
II-C	$\begin{bmatrix} 12.0810 \\ 1.3491 \\ 2.4587 \end{bmatrix}$	$\begin{bmatrix} 0.08 \\ 0.76 \end{bmatrix}$	$\begin{bmatrix} 0.05 \\ 0.95 \end{bmatrix}$	None	35.0000	11.9994

and $\Sigma + \Sigma_a$ for the new set of data are given in Tables 4-9 and 4-10. Ten samples of 1,500 subjects were generated for population I-C and twenty-five samples of 1,500 subjects were generated for population II-C. Populations I-C and II-C were chosen to have additional data generated in addition to the first set because these two cases are the most realistic.

Description of the Generation Routine

The present study requires that data be generated from a multivariate normal distribution with mean $\underline{\mu}$ and covariance matrix $\Sigma + \Sigma_a$, where the within covariance matrix (Σ) and the between covariance matrix (Σ_a) are specified as in Table 4-2. The generation procedure is composed of five steps:

1. Specify the values for the parameters so that they approximate the actual data. The Keesling and Wiley (1974) which analyzed real hierarchical data was used as a guide. This provided values for the pooled within-group regression coefficients ($\underline{\beta}$), the between-group regression coefficients for the individual level variables ($\underline{\beta}_a$), the regression coefficients for the group level variables ($\underline{\beta}_z$), the population means ($\underline{\mu}$), error variance defined at the individual level (σ^2), and at the group level (σ_a^2) as shown in Tables 4-3 and 4-4 for the first and second situation respectively. The population covariance matrices of the predictors were also specified based on the Keesling and Wiley study as shown in Table 4-5. The number of schools (k) and the number of subjects in each school (n) were specified a priori.

2. Compute the within and between covariance matrices (Σ and Σ_a)

Table 4-9

Population Covariance Matrices of the Predictor Variables of the Second Set of Data

Covariance Matrices	Population I-C	Population II-C
$\sum^{(x)}$	$\begin{bmatrix} 36.3347 & 4.7243 \\ 4.7423 & 61.4263 \end{bmatrix}$	$\begin{bmatrix} 81.0000 & 18.0000 \\ 18.0000 & 100.0000 \end{bmatrix}$
$\sum_a^{(x)}$	$\begin{bmatrix} 14.7149 & 2.9871 \\ 2.9871 & 25.8970 \end{bmatrix}$	$\begin{bmatrix} 35.0000 & 11.6383 \\ 11.6383 & 43.0000 \end{bmatrix}$
$\sum_a^{(z)}$	$\begin{bmatrix} 0.0072 & 0.0007 \\ 0.0007 & 0.0009 \end{bmatrix}$	Not Applicable
$\sum_a^{(xz)}$	$\begin{bmatrix} 0.0159 & 0.0065 \\ 0.0260 & 0.0088 \end{bmatrix}$	Not Applicable

Table 4-10

Population Covariance Matrices of the Second Set of Data

Case	Within Covariance (Σ)	Between Covariance (Σ_a)	Total Covariance ($\Sigma+\Sigma_a$)
I-C	346.746 (Symmetric) 93.439 36.335 31.609 4.724 61.426 0 0 0 0 0 0 0 0 0	82.094 (Symmetric) 24.074 14.715 27.505 2.987 25.897 0.077 0.216 0.026 0.007 0.022 0.007 0.009 0.001 0.001	428.840 (Symmetric) 117.513 51.050 59.114 47.711 87.323 0.077 0.016 0.026 0.007 0.022 0.007 0.009 0.001 0.001
II-C	95.467 (Symmetric) 20.160 81.000 77.440 18.000 100.000	52.000 (Symmetric) 12.806 35.000 41.432 11.638 43.000	147.467 (Symmetric) 32.966 116.000 118.872 29.638 143.000

between the outcome measure and the predictors as specified in Table 4-2.

3. Generate a random sample of k vectors $\underline{a}_i$, where $\underline{a}_i$ is multivariate normally distributed with mean vector $\underline{0}$ and covariance matrix Σ_a. A random sample of k vectors $\underline{a}_i$ are generated with the following procedure.

a. Generate 12 independent random variables which are uniformly distributed between zero and one. Software for the CDC 6500 has been developed which generates independent values of a random variable which is uniformly distributed over the range (0, 1), the values zero and one are excluded. This function, called Ranf, is described in Fortran reference manual version four (1978).

b. Convert the values from a uniform distribution to values from the normal distribution by Teichroew's method to approximate the inverse of the probability function for the standard normal distribution. Teichroew used a polynomial approximation to evaluate the inverse function. His procedure generates 12 independent random variables, $U_1, U_2, \ldots, U_{12}$, uniformly distributed between zero and one. Then R is defined as (Knuth, 1968):

$$R = (U_1 + U_2 + \ldots + U_{12} - 6)/4$$

The normal deviate, z is then approximated by:

$$z = ((((a_9 R^2 + a_7) R^2 + a_5) R^2 + a_3) R^2 + a_1) R$$

where $a_1 = 3.949846138$

$a_3 = 0.252408784$

$a_5 = 0.076542912$

$a_7 = 0.008355968$

$a_9 = 0.029899776$

For the first situation, each observation needed in this study consisted of 5 measures. Those 5 measures are the outcome variable (Y), two predictors defined at the individual level ($\underline{X}$), and two predictors defined only at the group level ($\underline{Z}$). For the second situation, each observation consists of 3 measures, the outcome variable (Y), and two predictors defined at the individual level ($\underline{X}$). Therefore, the procedure from a to b is repeated to obtain a 5x1 vector $\underline{z}$ for the first situation and a 3x1 vector $\underline{z}$ for the second situation which is normally distributed with a mean vector of zero and an identity matrix as the covariance matrix.

c. Transform $\underline{z}$ to $\underline{a}$ where $\underline{a}$ is normally distributed ($\underline{0}$, Σ_a). The transformation is:

$$\underline{a} = T\underline{z}$$

where T is the cholesky factor of Σ_a. The cholesky factor is a lower triangular matrix such that $TT' = \Sigma_a$. This is used because the covariance matrix of the transformed variables $\underline{a}$ is:

$$\text{Var}(\underline{a}) = T \text{ Var } (\underline{a})T'.$$

In this case, Var($\underline{a}$) is the identity matrix. Thus,

$$\text{Var}(\underline{a}) = TT' = \Sigma_a$$

which gives the desired result (Morrison, 1976). After the transformation, $\underline{a}$ is multivariate distributed normally with mean vector $\underline{0}$ and covariance matrix Σ_a.

4. Generate a random sample of kn vectors $\underline{e}_{ij}$ where $\underline{e}_{ij}$ is multivariate normally distributed with mean vector $\underline{0}$ and covariance matrix Σ. A random sample of kn vectors $\underline{e}_{ij}$ are generated with the same procedure as used in the generation of vector $\underline{a}_i$ except that here we generate kn vectors, and the covariance matrix is Σ instead of Σ_a.

5. Add the k values of $\underline{a}_i$ and kn values of $\underline{e}_{ij}$ to the μ according to formula (4-1) resulting in kn values of $\underline{W}_{ij}$. The values of $\underline{a}_i$ are constant for the ith group, i.e.,

$$(4\text{-}1) \quad \underline{W}_{ij} = \underline{\mu} + \underline{a}_i + \underline{e}_{ij}$$

where $\underline{W} = \begin{matrix} Y \\ \underline{X} \end{matrix}$ for the first situation

and $\underline{W} = \begin{matrix} \underline{Z} \\ Y \\ \underline{X} \end{matrix}$ for the second situation.

The program MYDATA (see appendix A) was written for this study to generate a random sample of kn vectors of $\underline{W}_{ij}$ where $\underline{W}_{ij}$ is multivariate normally distributed with mean vector $\underline{\mu}$ and covariance matrix $\Sigma + \Sigma_a$, using the procedure described above.

For each sample the pooled within and between mean square matrices (S and S_a) are computed as shown in formulas (4-2) and (4-3) respectively:

$$(4\text{-}2) \quad S = \frac{1}{k(n-1)} \sum_{i}^{k} \sum_{j}^{n} (\underline{W}_{ij} - \underline{W}_i)(\underline{W}_{ij} - \underline{W}_i)'$$

where the expected value of S is the pooulation within covariance matrix and the $E(S) = \Sigma$ and

$$(4\text{-}3) \quad S_a = \frac{1}{k-1} n \sum_{i}^{k} (\underline{W}_i - \underline{W})(\underline{W}_i - \underline{W})'$$

where the expected value of S_a is the following:

$$E(S_a) = \Sigma + n\Sigma_a$$

Here, Σ_a is the population between levels covariance matrix. The general structure of S is the following:

$$S = \begin{vmatrix} S_y & S_{yx} \\ S_{xy} & S_x \end{vmatrix}$$

where S_y is the pooled within variance of Y, S_{xy} is the pooled within covariance matrix between X and Y, S_x is the pooled within covariance matrix of $\underline{X}$.

To compute an estimate of the pooled within-group regression coefficient ($\underline{\beta}$) for any approach the formula (4-4) is used.

$$(4\text{-}4) \quad \underline{\beta} = S_x^{-1} S_{xy}$$

For the first situation where there are both individual level predictors and group level predictors, four approaches were investigated: two stage analysis, group level analysis, full model analysis, and the subtraction approach. The main concern is to estimate the regrassion coefficients for the group level variables ($\underline{\beta}_z$) by those four approaches.

For the second situation where there were only individual level predictors, four approaches were investigated: group level analysis, Bock application, subtraction analysis and full model analysis. The main purpose of each approach is to estimate the between group regression coefficients for the individual level variables ($\underline{\beta}_a$). The procedure for each analysis approach is described in following sections.

Two Stage Analysis Approach

The procedure to estimate $\underline{\beta}_z$ by using the two stage analysis approach is the following:

1. Compute an estimate of the pooled within-group regression coefficient ($\underline{\beta}$) using formula (4-4).

2. Compute an estimate of the group mean ($\hat{\mu}_{y_i}$) using formula (4-5).

$$(4\text{-}5) \quad \hat{\mu}_{y_i} = \hat{\mu}_y + \hat{\underline{\beta}}'(\hat{\underline{\mu}}_{x_i} - \hat{\underline{\mu}}_x)$$

3. Compute $\underline{\beta}_z$ using equation (4-6) implemented by the Finn multivariance program (1972).

$$(4\text{-}6) \quad \mu_{y_i} = \mu_y + \underline{\beta}_z'(\underline{Z}_i - \underline{\mu}_z) + \lambda\hat{\mu}_{y_i} + \delta_i$$

Group Level Analysis Approach

Under the group level analysis approach the $\underline{\beta}_z$ for the first situation and $\underline{\beta}_a$ for the second situation are estimated separately from $\underline{\beta}$. The Finn multivariance program (1972) is used to estimate $\underline{\beta}_z$ under equation (4-7) and $\underline{\beta}_a$ under the equation (4-8).

$$(4\text{-}7) \quad \mu_{y_i} = \mu_y + \underline{\beta}_z'(\underline{Z}_i - \underline{\mu}_z) + \underline{\beta}_a'(\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i$$

$$(4\text{-}8) \quad \mu_{y_i} = \mu_y + \underline{\beta}_a'(\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i$$

Subtraction Analysis Approach

For the first situation, $\underline{Z}$ variables are defined only at the group level. The procedure of estimating $\underline{\beta}_z$ by the subtraction approach is the same as for the group level analysis approach. The Finn multivariance program is used to estimate $\underline{\beta}_z$ under equation (4-7).

To obtain the correct estimates of $\underline{\beta}_a$ in the second situation Keesling recommends performing three steps as follows.

1. Compute estimates of the pooled within-group regression coefficient ($\underline{\beta}$) using formula (4-4). This step is to compute $\underline{\beta}$ under

the model of (4-9):

$$(4\text{-}9) \quad Y_{ij} = \mu_{y_i} + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_{x_i}) + \varepsilon_{ij}$$

2. Compute the estimates of the between-group regression coefficient ($\underline{\beta}_a^*$) with equation (4-10) using the Finn multivariance program.

$$(4\text{-}10) \quad \mu_{y_i} = \mu_y + \underline{\beta}_a^{*\prime}(\underline{\mu}_{x_i} - \underline{\mu}_x) + \delta_i$$

3. Compute the correct estimates of the between-group regression coefficients for the individual level variables ($\underline{\beta}_a$) by using formula (4-11).

$$(4\text{-}11) \quad \underline{\beta}_a = \underline{\beta}_a^* - \underline{\beta}$$

Full Model Analysis Approach

The full model analysis approach used the individual outcome as the dependent variable, the individual level variables, the mean of the individual variables and the variables that are defined at the group level as the predictors. The Finn multivariance program is used to estimate $\underline{\beta}_z$ for the first situation under equation (4-12) and $\underline{\beta}_a$ for the second situation under equation (4-13).

$$(4\text{-}12) \quad Y_{ij} = \mu_y + \underline{\beta}_z'(\underline{Z}_i - \underline{\mu}_z) + \underline{\beta}_a'(\underline{\mu}_{x_i} - \underline{\mu}_x) + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_x) + \varepsilon_{ij}$$

$$(4\text{-}13) \quad Y_{ij} = \mu_y + \underline{\beta}_a'(\underline{\mu}_{x_i} - \underline{\mu}_x) + \underline{\beta}'(\underline{X}_{ij} - \underline{\mu}_x) + \varepsilon_{ij}$$

Bock Application Analysis Approach

Bock's analysis approach provides an estimate of the between covariance matrix (Σ_a) which is guaranteed to be at least a positive semi-definite covariance matrix, and then from this matrix estimates the regression coefficient. The steps to this approach are as follows:

1. Use the Finn multivariance program to determine discrimination function coefficients $\underline{t}_1$, and canonical variances ϕ_1 (1=1, 2, 3).

2. Compute the positive semi-definite between covariance matrix (Σ_a) using formula (4-14).

$$(4\text{-}14) \quad \hat{\Sigma}_a = [(T)^{-1}{}'(\Phi - I)T^{-1}]/n$$

where elements in the columns of T are the discrimination function coefficients $\underline{t}_1$, and the diagonal elements of diagonal matrix Φ are significant canonical variances Φ_1 (1=1, 2, . . . , s) and p-s unities (p is the dimension of T and s is less than and equal to p). When all canonical variance ϕ are significant (s=p)

$$\hat{\Sigma}_a = [S_a - S]/n$$

where S and S_a are within and between mean square matrices that are computed by formula (4-2) and (4-3) respectively.

3. Using Σ_a to estimate $\underline{\beta}_a$ by formula (4-15).

$$(4\text{-}15) \quad \hat{\underline{\beta}}_a = \hat{\Sigma}_a^{(x)-1}\hat{\Sigma}_a^{(xy)}$$

The general structure of $\hat{\Sigma}_a$ is the following:

$$\hat{\Sigma}_a = \begin{bmatrix} \hat{\Sigma}_a^{(y)} & \hat{\Sigma}_a^{(yx)} \\ \hat{\Sigma}_a^{(xy)} & \hat{\Sigma}_a^{(x)} \end{bmatrix}$$

where $\hat{\Sigma}_a^{(y)}$ is the between variation of Y, $\hat{\Sigma}_a^{(xy)}$ is the between covariance of $\underline{X}$ and Y, and $\hat{\Sigma}_a^{(x)}$ is the between covariance of $\underline{X}$.

CHAPTER V

SIMULATION RESULTS

The simulation procedures employed in this study to investigate the methods of analyzing hierarchical data were reviewed in Chapter IV. The main purpose of this dissertation is to determine which approach gives the best estimates of the between and within-group regression coefficients in terms of accuracy (least amount of bias) and in terms of precision for various situations. Six populations as shown in Table 4-1 were used as the basis from which the questions of interest were explored. For each population, 50 samples of size 1,500 were generated. In each sample, there were 50 schools with 30 students nested within each school. In order to confirm the emerging conclusions resulting from the analyses, 10 additional samples for population I-C and 25 additional samples for population II-C with the new set of parameter values as given in Table 4-8, were generated. Data from population I-A, I-B and I-C were analyzed by the two stage analysis approach (as suggested by Keesling and Wiley), group level analysis approach (as suggested by Cronbach and Webb), subtraction analysis approach (as suggested by Keesling) and full model analysis approach (as suggested by Keesling). Data from population II-A, II-B, and II-C were analyzed by Bock application analysis approach, group level analysis approach, subtraction analysis approach and full model analysis approach.

The results of the data analysis of population I-A are shown in

Table 5-1. All four approaches give good estimates of the pooled within-group regression coefficients β_1 and β_2. The means of the estimates of $\hat{\beta}_1$ and $\hat{\beta}_2$ over the 50 samples are 2.524 and 0.322 for all four approaches while the values of the parameters of β_1 and β_2 are 2.530 and 0.320. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are 0.009 and 0.002 for all four approaches. The result of testing the hypotheses that the mean of the estimates of the within-group regression coefficients of all 50 samples are equal to the parameters β_1 and β_2 are not significant at the 0.01 level ($t = -0.667$, $t = 1.000$). The ratios of the bias squared to mean square error which are computed by formula (5-1) for $\hat{\beta}_1$ and $\hat{\beta}_2$ are 0.008 and 0.028. The formula is:

$$(5\text{-}1) \quad \text{Bias ratio} = \frac{(\text{Bias})^2}{\text{mean square error}}$$

where Bias = average value of the estimates - parameter value and mean square error = variance of estimator + $(\text{bias})^2$. We can conclude that all four approaches give good estimates of β_1 and β_2 since the bias ratios are quite small. The results of the hypothesis tests showed that the means of $\hat{\beta}_1$ and $\hat{\beta}_2$ over the 50 samples are not different from the values of the parameters β_1 and β_2.

The means of the estimates of β_{z1} and β_{z2} analyzed by the two stage analysis approach are closer to the parameters than the other three analysis approaches ($\bar{\hat{\beta}}_{z1} = 3.905$, $\bar{\hat{\beta}}_{z2} = 1.722$, $\beta_{z1} = 4.08$, $\beta_{z2} = 2.15$). The results of testing the hypotheses that the mean of the estimates of the regression coefficients defined for the group level variables over all 50 samples are equal to the parameters β_{z1} and β_{z2} are not significant for the two stage analysis approach, whereas, the tests for the other three analysis approaches are significant at the

Table 5-1

Simulation Results of Population I-A

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Two stage[1]	2.524	0.066	0.009	-0.667	0.008
β_1= 2.53	Group level	2.524	0.066	0.009	-0.667	0.008
	Full model	2.524	0.066	0.009	-0.667	0.008
	Subtraction	2.524	0.066	0.009	-0.667	0.008
	Two stage	0.322	0.012	0.002	1.000	0.028
β_2= 0.32	Group level	0.322	0.012	0.002	1.000	0.028
	Full model	0.322	0.012	0.002	1.000	0.028
	Subtraction	0.322	0.012	0.002	1.000	0.028
	Two stage	3.905	0.920	0.130	-1.346	0.036
β_{z1}=4.08	Group level	3.651	0.858	0.121	-3.545*	0.203
	Full model	3.651	0.858	0.121	-3.545*	0.203
	Subtraction	3.651	0.858	0.121	-3.545*	0.203
	Two stage	1.722	1.795	0.254	-1.682	0.055
β_{z2}=2.15	Group level	1.405	1.709	0.242	-3.079*	0.162
	Full model	1.405	1.709	0.242	-3.079	0.162
	Subtraction	1.405	1.709	0.242	-3.079	0.162

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

[1]All four approaches gave exactly the same estimates of β_1 and β_2 as was expected.

0.01 level of significance. The bias ratios of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ analyzed by the two stage analysis are 0.036 and 0.055 while the bias ratios for the other three approaches are 0.203 and 0.162. We can conclude that in the situation where there is no group level effect ($\underline{\beta}_a = \underline{0}$) two stage analysis approach gives the best estimates of the regression coefficients defined for the group level variables. The empirical sampling distributions of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ over the 50 samples are shown in Figures 5-1 and 5-2. According to Figures 5-1 and 5-2, all four approaches have very similar distributions.

The results of the data analysis of population I-B are shown in Table 5-2. The parameter values of β_1 and β_2 are 2.53 and 0.32, respectively. All four approaches gave the same average estimates for β_1 and β_2 (2.519 and 0.322). The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ for all four approaches are quite small, 0.011 and 0.001. The bias ratio of $\hat{\beta}_1$ and $\hat{\beta}_2$ for all four approaches are 0.019 and 0.033. The result of testing the hypotheses that the mean of the estimates for the within-group regression coefficients over all 50 samples are equal to the parameters β_1 and β_2 are not significant at the 0.01 level of significance ($t = -1.000$, $t = 2.000$). Therefore, the results of the data analysis of population I-B indicate that all four approaches gave good estimates of the within-group regression coefficients with good precision and small bias ratios.

The means of the estimates of β_{z1} and β_{z2} analyzed by the four approaches are almost the same. However, the two stage analysis approach gave a somewhat better estimate than the other three approaches as demonstrated by the fact that the bias ratios of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ are smaller. The bias ratio of the estimates of β_{z1} and β_{z2} are 0.005 and 0.018 for

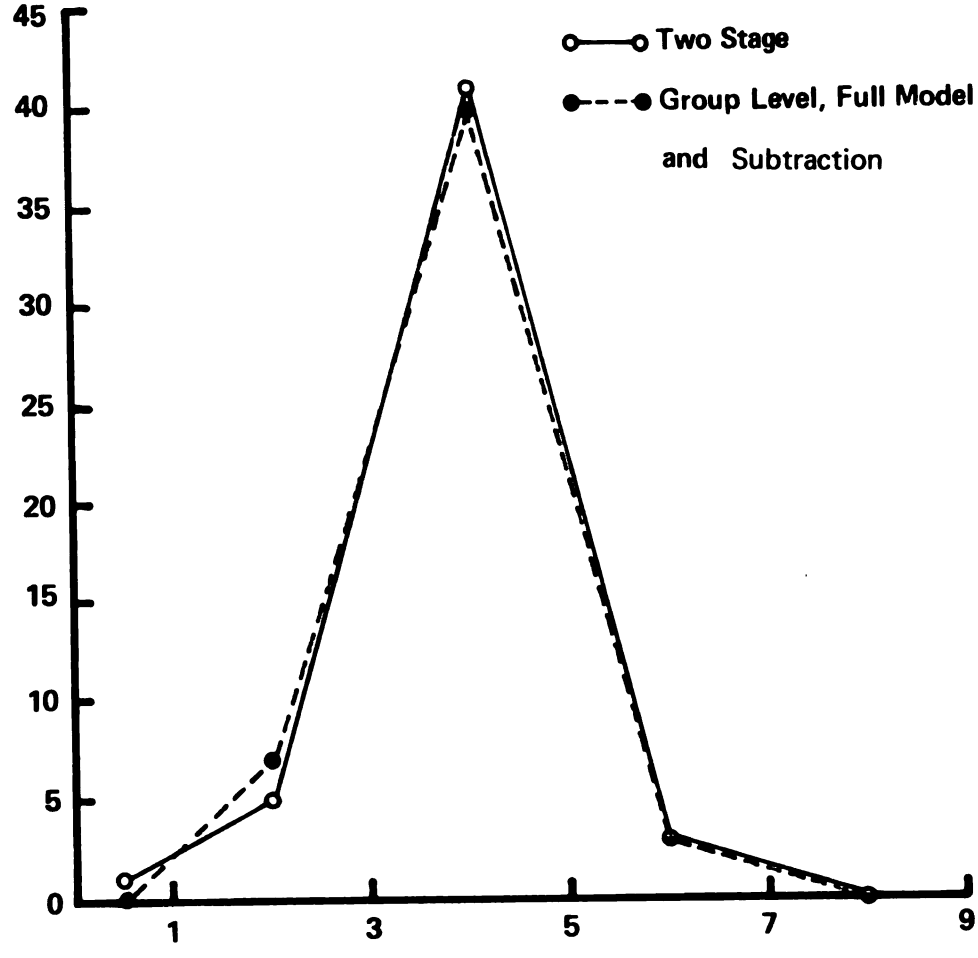

Figure 5-1 Sampling Distribution of β_{z1} From Population I-A

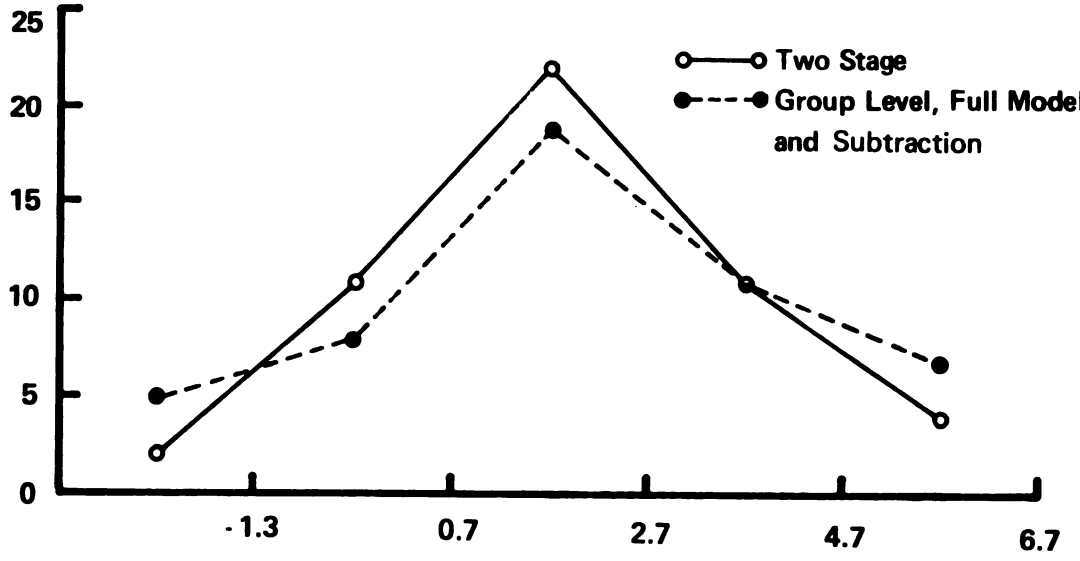

Figure 5-2 Sampling Distribution of β_{z2} From Population I-A

Table 5-2

Simulation Results of Population I-B

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Two stage	2.519	0.080	0.011	-1.000	0.019
$\beta_1 = 2.53$	Group level	2.519	0.080	0.011	-1.000	0.019
	Full model	2.519	0.080	0.011	-1.000	0.019
	Subtraction	2.519	0.080	0.011	-1.000	0.019
	Two stage	0.322	0.011	0.001	2.000	0.033
$\beta_2 = 0.32$	Group level	0.322	0.011	0.001	2.000	0.033
	Full model	0.322	0.011	0.001	2.000	0.033
	Subtraction	0.322	0.011	0.001	2.000	0.033
	Two stage	4.203	1.782	0,252	0.488	0.005
$\beta_{z1}=4.08$	Group level	4.236	1.907	0.270	0.578	0.007
	Full model	4.235	1.907	0.270	0.578	0.007
	Subtraction	4.235	1.907	0.270	0.578	0.007
	Two stage	1.381	5.651	0.799	-0.962	0.018
$\beta_{z2}=2.15$	Group level	1.335	5.901	0.834	-0.977	0.019
	Full model	1.330	5.901	0.834	-0.977	0.019
	Subtraction	1.335	5.901	0.834	-0.977	0.019

**Ratio of the estimate of the bias to mean square error.

the two stage analysis approach, and 0.007 and 0.019 for the other three approaches. The results of testing the hypotheses that the means of the estimates of the regression coefficients defined for the group level variables over all 50 samples are equal to the parameters β_{z1} and β_{z2} are not significant at the 0.01 level of significance for all four approaches. We conclude that in the situation where the group level effects are equal to the individual effects ($\underline{\beta}_a = \underline{\beta}$) all four approaches give good estimates of the regression coefficients defined for the group level variables. The sampling distributions of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ are shown in Figures 5-3 and 5-4.

The results of the data analysis of population I-C are shown in Table 5-3. The parameter values of β_1 and β_2 are 2.53 and 0.32. The means of the estimates of β_1 and β_2 are 2.512 and 0.321 for all four approaches. The standard errors for all four approaches are quite small (0.010 and 0.002). The bias ratios of β_1 and β_2 for all four approaches are 0.067 and 0.005. The results of testing the hypotheses that the means of the estimates of within-group regression coefficient of all 50 samples are equal to the parameters β_1 and β_2 are not significant at the 0.01 level of significance ($t = 0.067$, $t = 0.005$). So, all four approaches gave the same good estimates of within-group regression coefficients with high precision and small bias ratios.

The means of the estimates of the regression coefficients defined for the group level variables ($\bar{\hat{\beta}}_{z1}$ and $\bar{\hat{\beta}}_{z2}$) from the parameter values than in the other three approaches. The results of testing the hypotheses that the means of the estimates of the regression coefficients defined for the group level variables over all 50 samples are equal to the parameters β_{z1} and β_{z2} are significant at the 0.01 level of

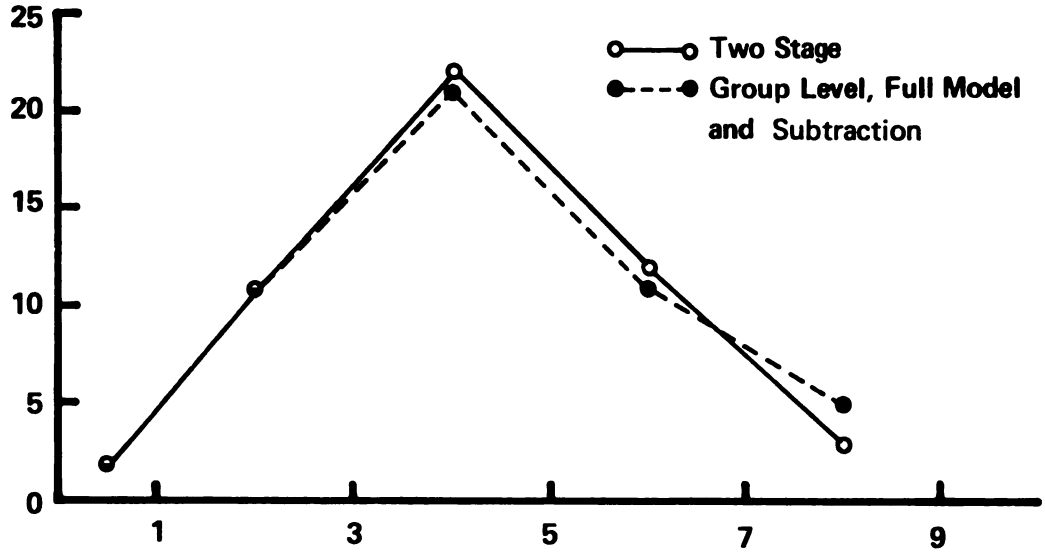

Figure 5-3 Sampling Distribution of β_{z1} From Population I-B

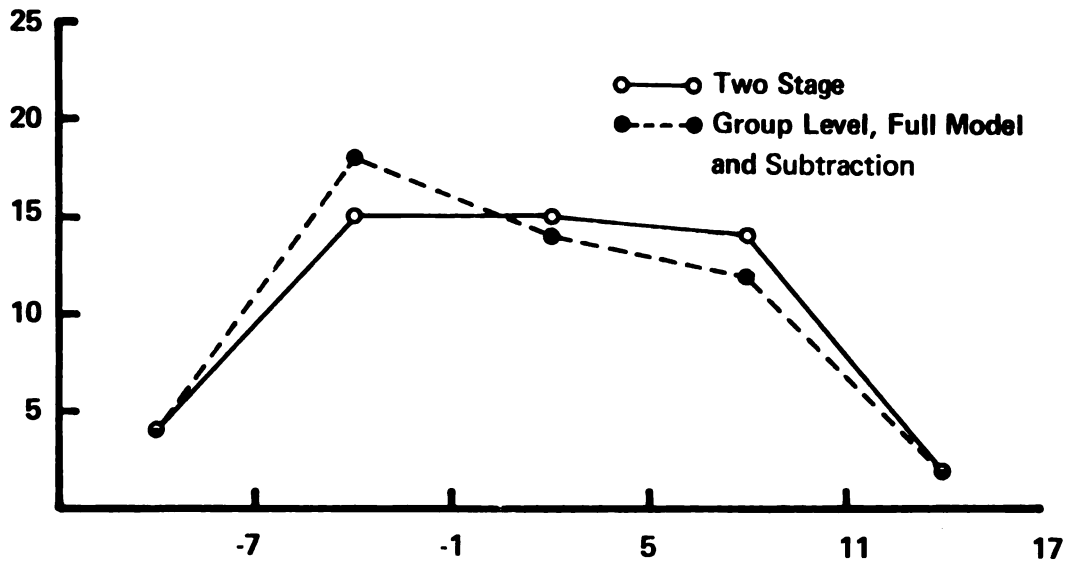

Figure 5-4 Sampling Distribution of β_{z2} From Population I-B

Table 5-3

Simulation Results of Population I-C

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Two stage	2.512	0.068	0.010	-1.800	0.067
$\beta_1 = 1.53$	Group level	2.512	0.068	0.010	-1.800	0.067
	Full model	2.512	0.068	0.010	-1.800	0.067
	Subtraction	2.512	0.068	0.010	-1.800	0.067
	Two stage	0.321	0.014	0.002	0.500	0.005
$\beta_2 = 0.32$	Group level	0.321	0.014	0.002	0.500	0.005
	Full model	0.321	0.014	0.002	0.500	0.005
	Subtraction	0.321	0.014	0.002	0.500	0.005
	Two stage	5.325	1.986	0.281	4.527	0.286
$\beta_{z1} = 4.08$	Group level	4.296	1.946	0.275	0.785	0.012
	Full model	4.296	1.946	0.275	0.785	0.012
	Subtraction	4.926	1.946	0.275	0.785	0.012
	Two stage	4.932	4.937	0.697	3.991*	0.245
$\beta_{z2} = 2.15$	Group level	3.194	4.911	0.695	1.502	0.044
	Full model	3.194	4.911	0.695	1.502	0.044
	Subtraction	3.194	4.911	0.695	1.502	0.044

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

significance for the other three approaches ($t = 3.062$, $t = 3.991$). These were not significant at the 0.01 level of significance for the other three approaches ($t = 0.785$, $t = 1.502$). The bias ratios of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ analyzed by the other three approaches are 0.012 and 0.044. We can conclude that in the situation where the group level effects are not equal to the individual level effects ($\underline{\beta}_a \neq \underline{\beta} \neq \underline{0}$), the group level analysis approach, the subtraction analysis approach and the full model analysis approach gave the better estimates of the regression coefficients defined for the group level variables as opposed to the two stage analysis approach. The sampling distributions of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ are shown in Figures 5-5 and 5-6.

In the situation where there were only individual level explanatory variables, the data for three populations II-A, II-B, and II-C were analyzed by the group level analysis approach, the Bock application approach, the full model analysis approach, and the subtraction analysis approach. The results of the data analysis of population II-A are shown in Table 5-4. All four approaches gave the same estimates of the pooled within-group regression coefficients. The means of the estimates of β_1 and β_2 are 2.533 and 0.321 while the parameter values of β_1 and β_2 are 2.530 and 0.320 respectively. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ are 0.010 and 0.002 for all four approaches. The results of testing the hypotheses that the means of the estimates of within regression coefficients of all 50 samples are equal to the parameters β_1 and β_2 are not significant at the 0.01 level of significance ($t = 0.300$, $t = 0.500$) for all four approaches. The bias ratios of $\hat{\beta}_1$ and $\hat{\beta}_2$ are 0.002 and 0.004. All four approaches yielded the same estimates of the within-group regression coefficients with high precision

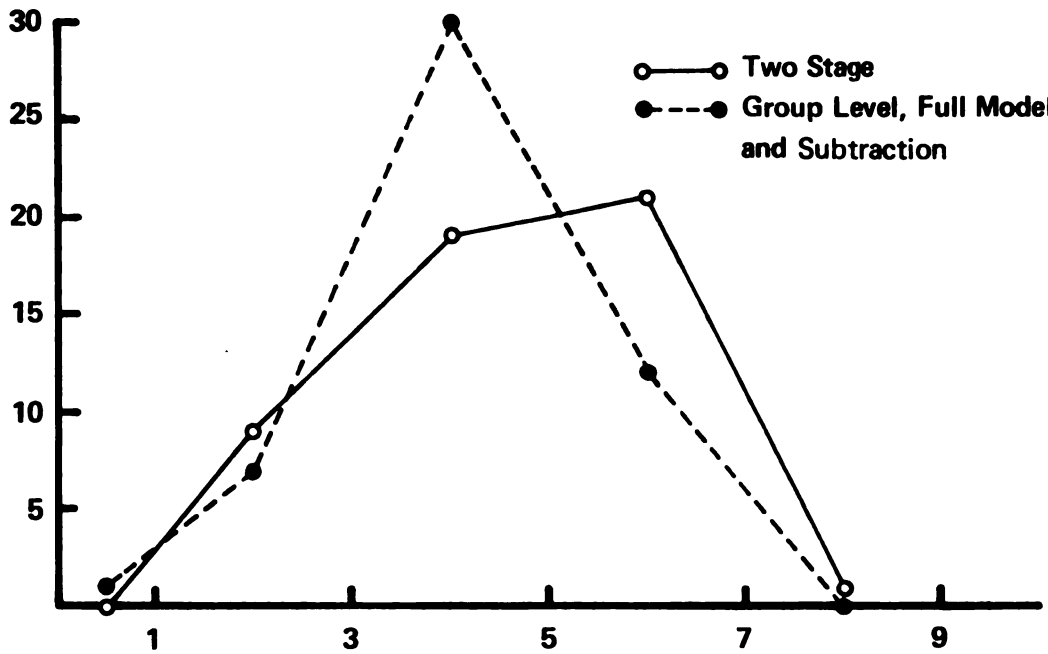

Figure 5-5 Sampling Distribution of β_{z1} From Population I-C

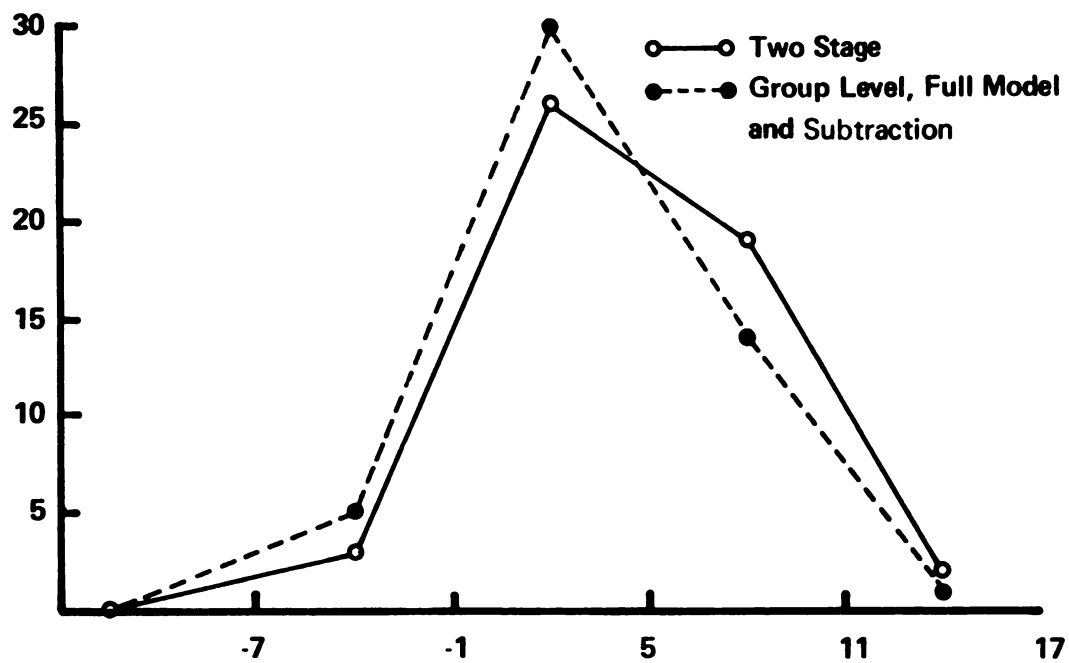

Figure 5-6 Sampling Distribution of β_{z2} From Population I-C

Table 5-4

Simulation Results of Population II-A

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Group level	2.533	0.070	0.010	0.300	0.002
$\beta_1 = 2.53$	Bock application	2.533	0.070	0.010	0.300	0.002
	Full model	2.533	0.070	0.010	0.300	0.002
	Subtraction	2.533	0.070	0.010	0.300	0.002
	Group level	0.321	0.015	0.002	0.500	0.004
$\beta_2 = 0.32$	Bock application	0.321	0.015	0.002	0.500	0.004
	Full model	0.321	0.015	0.002	0.500	0.004
	Subtraction	0.321	0.015	0.002	0.500	0.004
	Group level	0.017	0.146	0.021	0.810	0.013
$\beta_{a1} = 0.00$	Bock application	0.008	0.148	0.021	0.381	0.003
	Full model	-2.519	0.168	0.024	-104.958*	0.996
	Subtraction	-2.159	0.168	0.024	-104.958*	0.996
	Group level	-0.011	0.161	0.023	-0.478	0.002
$\beta_{a2} = 0.00$	Bock application	-0.067	0.181	0.026	-2.577	0.123
	Full model	-0.338	0.168	0.024	-14.083*	0.805
	Subtraction	-0.338	0.168	0.024	-14.083*	0.805

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

and small bias ratios.

The means of the estimates of the between-group regression coefficients for the individual level variables ($\bar{\hat{\beta}}_{a1}$ and $\bar{\hat{\beta}}_{a2}$) analyzed by the group level analysis approach and the Bock application approach are similar and closer to the parameters ($\beta_{a1} = 0, \beta_{a2} = 0$). The full model analysis approach and the subtraction analysis approach gave the same estimates of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ and they are not close to the parameter values ($\hat{\beta}_{a1} = -2.519, \hat{\beta}_{a2} = -0.338$). The results of testing the hypothesis that the means of the estimates of the between-group regression coefficients for the individual level variables over the 50 samples are equal to the parameters β_{a1} and β_{a2} are not significant for the group level analysis approach and the Bock application analysis approach. However, they are significant at 0.01 level of significance for the full model analysis approach and the subtraction analysis approach. The bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the group level analysis approach and the Bock application analysis approach were quite small (the bias ratios of $\hat{\beta}_{a1}$ for the group level analysis approach and the Bock application analysis approach were 0.013 and 0.003, the bias ratio of $\hat{\beta}_{a2}$ for the group level analysis approach and the Bock application analysis approach were 0.002 and 0.123), while the bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the other analysis approaches were quite large (bias ratio of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for these two approaches were 0.996 and 0.805). We can conclude that in the situation where there was no group level effect ($\beta_{a1} = \beta_{a2} = 0$), the group level analysis approach and the Bock application analysis approach gave the better estimates of the between-group regression coefficients while the full model and subtraction analysis approach gave incorrect estimates of the between-group regression coefficients. The sampling distributions

of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ are shown in Figures 5-7 and 5-8.

The results of the data analyses for population II-B are shown in Table 5-5. All four approaches gave the same estimates of the pooled within regression coefficients. The means of the estimates of β_1 and β_2 were 2.546 and 0.320. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ were 0.010 and 0.002 for all four approaches. The results of testing the hypothesis that the means of the estimates of the pooled within regression coefficients over all 50 samples were equal to the parameters β_1 and β_2 were not significant at 0.01 level of significance ($t = 0.000$, $t = 0.000$). The bias ratios of $\hat{\beta}_1$ and $\hat{\beta}_2$ were 0.051 and 0.000. All four approaches gave good estimates of the pooled within regression coefficients.

The means of the estimates of the between-group regression coefficients for the individual level variables analyzed by the group level analysis approach and the Bock application approach were quite similar and close to the paremeter values. The means of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the group level analysis approach were 2.525 and 0.310, and for the Bock application analysis approach were 2.524 and 0.309. The full model analysis approach and the subtraction analysis approach gave the same estimates for β_{a1} and β_{a2} and they were not close to the parameter values ($\hat{\beta}_{a1} = -0.021$, $\hat{\beta}_{a2} = -0.014$). The results of testing the hypothesis that the means of the estimates of the between-group regression coefficients were equal to the parameters β_{a1} and β_{a2} were not significant for the group level analysis approach or for the Bock application analysis approach. However, they were significant at 0.01 level of significance when analyzed by the full model and subtraction analysis approaches. The bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the group level and

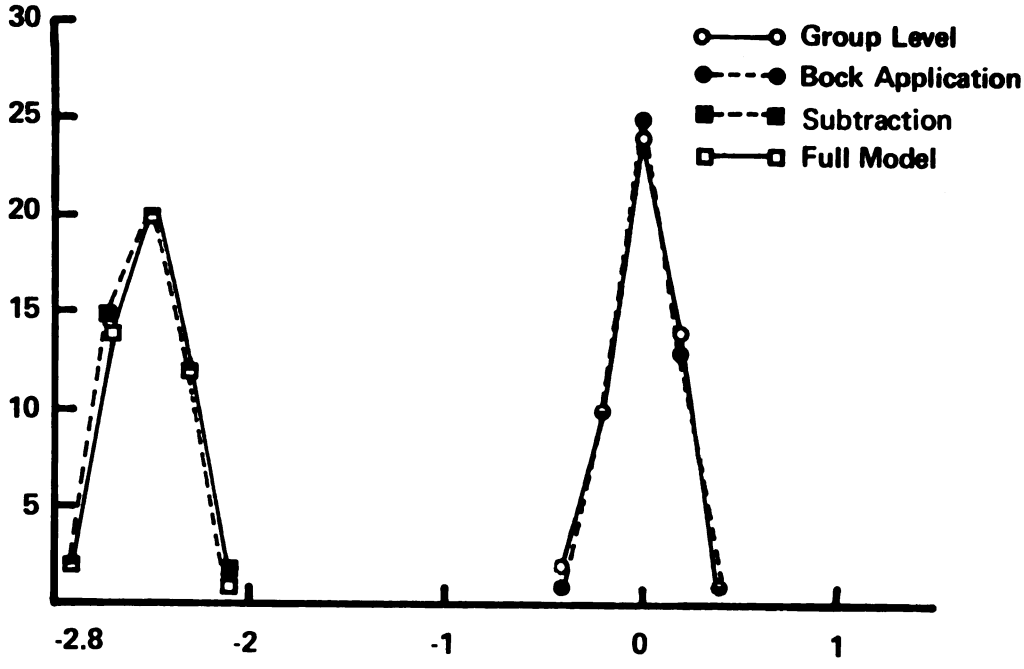

Figure 5-7 Sampling Distribution of β_{a1} From Population II-A

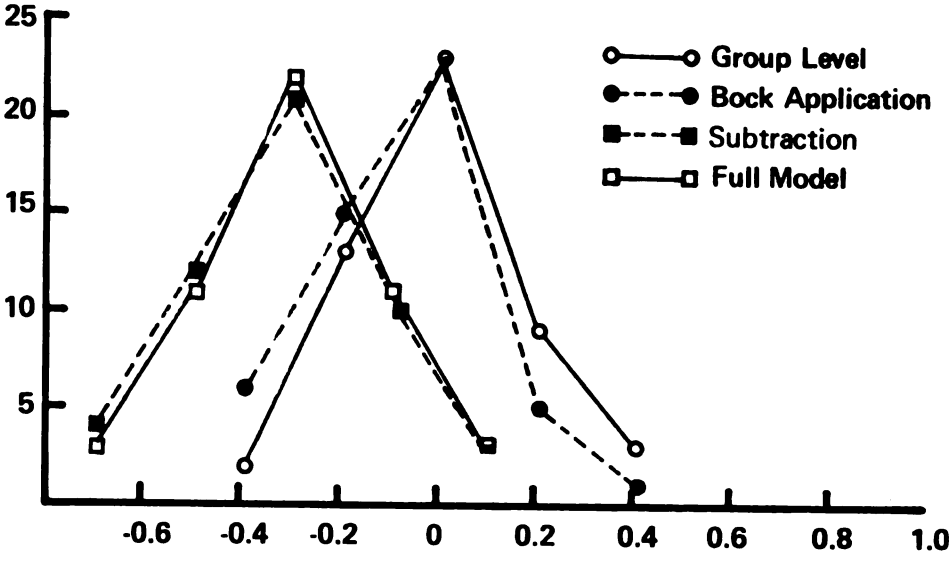

Figure 5-8 Sampling Distribution of β_{a2} From Population II-A

Table 5-5

Simulation Results of Population II-B

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Group level	2.546	0.070	0.010	0.600	0.051
$\beta_1 = 2.53$	Bock application	2.546	0.070	0.010	0.600	0.051
	Full model	2.546	0.070	0.010	0.600	0.051
	Subtraction	2.546	0.070	0.010	0.600	0.051
	Group level	0.320	0.012	0.002	0.000	0.000
$\beta_2 = 0.32$	Bock application	0.320	0.012	0.002	0.000	0.000
	Full model	0.320	0.013	0.002	0.000	0.000
	Subtraction	0.320	0.012	0.010	0.000	0.000
	Group level	2.525	0.221	0.031	-0.161	0.001
$\beta_{a1} = 2.53$	Bock application	2.524	0.222	0.031	-0.194	0.001
	Full model	-0.021	0.228	0.032	-79.719*	0.992
	Subtraction	-0.021	0.228	0.032	-79.719*	0.992
	Group level	0.310	0.158	0.022	-0.455	0.004
$\beta_{a2} = 0.32$	Bock application	0.309	0.178	0.025	-0.440	0.004
	Full model	-0.010	0.157	0.022	-15.000*	0.818
	Subtraction	-0.010	0.157	0.022	-15.000*	0.818

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

Bock application analysis approaches were quite small (bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for both the group level analysis approach and the Bock application analysis approach were 0.001 and 0.004) while the bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ analyzed by the other two analysis approaches were quite large (bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ equal to 0.992 and 0.822). We can conclude that in the situation where the between-group regression coefficients ($\underline{\beta}_a = \underline{\beta} \neq 0$), the group level and Bock application analysis approaches gave correct estimates of the between-group regression coefficients while the full model and subtraction analysis approaches gave incorrect estimates of the between-group regression coefficients. The sampling distributions of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ are shown in Figures 5-9 and 5-10.

The results of the data analysis of population II-C are shown in Table 5-6. All four approaches gave the same estimates of the pooled within regression coefficients. The means of the estimates of β_1 and β_2 were 2.518 and 0.323 while the standard errors were 0.010 and 0.002 for all four approaches. The results of testing the hypothesis was that the means of the estimates of the within regression coefficients were equal to the parameters were not significant ($t = -1.200$, $t = 1.500$). The bias ratios of $\hat{\beta}_1$ and $\hat{\beta}_2$ were 0.028 and 0.060 for all four approaches.

The means of the estimates of the between-group regression coefficients for the group level and Bock application analysis approaches were quite similar and close to the parameter values. The means of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ analyzed by group level analysis approach were 1.424 and 0.934 and for the Bock application analysis approach they were 1.412 and 0.946 ($\beta_{a1} = 1.45$ and $\beta_{a2} = 0.89$). The full model analysis

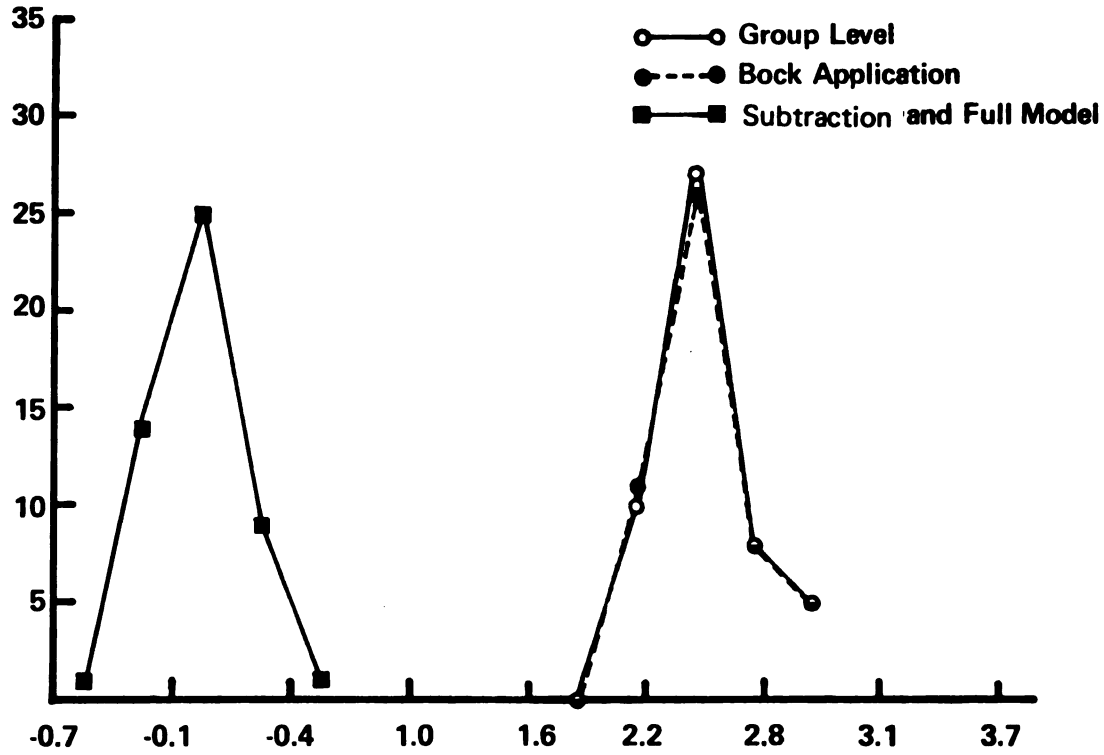

Figure 5-9 Sampling Distribution of β_{a1} From Population II-B

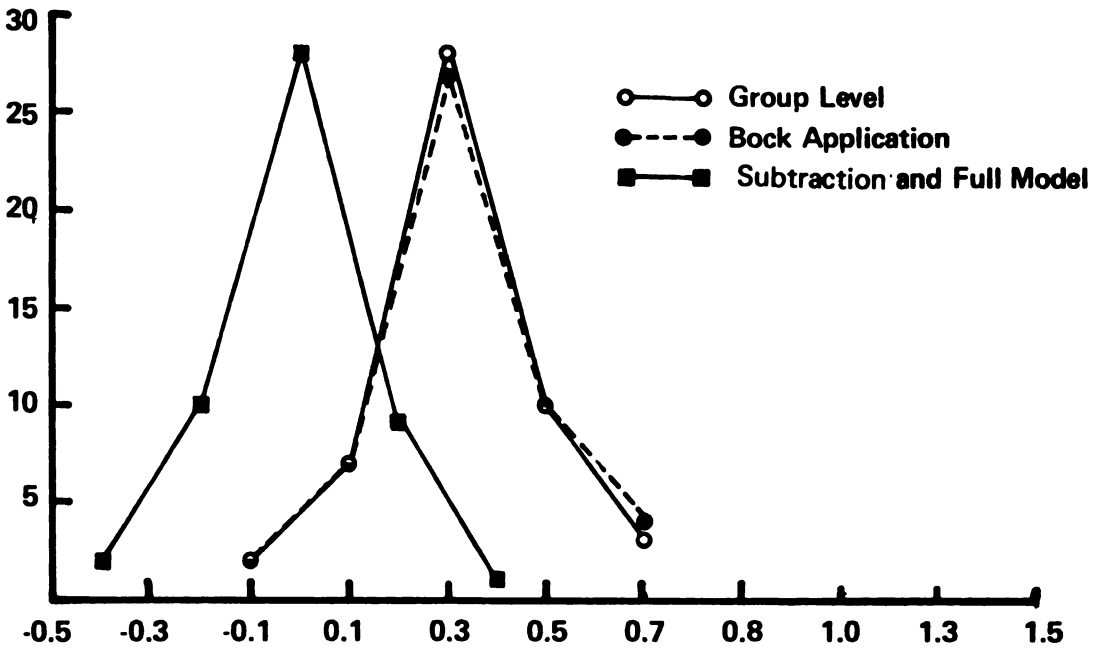

Figure 5-10 Sampling Distribution of β_{a2} From Population II-B

Table 5-6

Simulation Results of Population II-C

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Group level	2.518	0.071	0.010	-1.200	0.028
$\beta_1 = 2.53$	Bock application	2.518	0.071	0.010	-1.200	0.028
	Full model	2.518	0.071	0.010	-1.200	0.028
	Subtraction	2.518	0.071	0.010	-1.200	0.028
	Group level	0.323	0.012	0.002	1.500	0.060
$\beta_2 = 0.32$	Bock application	0.323	0.012	0.002	1.500	0.060
	Full model	0.323	0.012	0.002	1.500	0.060
	Subtraction	0.323	0.012	0.002	1.500	0.060
	Group level	1.424	0.260	0.037	-0.703	0.010
$\beta_{a1} = 1.45$	Bock application	1.412	0.184	0.026	-1.462	0.042
	Full model	-1.093	0.179	0.025	-101.720*	0.995
	Subtraction	-1.093	0.117	0.025	-101.720*	0.995
	Group level	0.934	0.196	0.021	2.095	0.049
$\beta_{a2} = 0.89$	Bock application	0.946	0.165	0.023	2.453	0.105
	Full model	0.611	0.021	0.021	-13.286*	0.782
	Subtraction	0.611	0.149	0.021	-13.286*	0.782

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

approach and the subtraction analysis approach yielded the same estimates for $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ were not close to the parameter values ($\bar{\hat{\beta}}_{a1}$ = 1.093, $\bar{\hat{\beta}}_{a2}$ = 0.611). In testing the hypothesis that the means of the estimates of the between-group regression coefficients were equal to the parameters β_{a1} and β_{a2} were not significant for the group level analysis approach and the Bock application analysis approach; however, they were significant for the full model and subtraction analysis approaches. The bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the group level analysis approach and the Bock application analysis approach were quite small while the same bias ratios for the other two analysis approaches were quite large. We can conclude that in the situation where the between-group regression coefficients were not equal to the within-group regression coefficients ($\underline{\beta}_a \neq \underline{\beta} \neq 0$), that the group level and Bock application analysis approaches gave the correct estimates of the between-group regression coefficients while the full model and the subtraction analysis approaches gave incorrect estimates of the between-group regression coefficients. The sampling distributions of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ are shown in Figures 5-11 and 5-12.

The specifications concerning $\underline{\beta}_a$ and $\underline{\beta}$ for populations I-C and II-C are most similar to the types of situations encountered in the real world. The intraclass correlations of variables Y and $\underline{X}$ are, however, quite high for both cases (0.90). Therefore, in order to check whether the analysis approaches give the same results for situations where the intraclass correlations are not as high as for the data presented in the preceding pages, a second set of data for populations I-C and II-C was generated with new parameter values chosen so that the intraclass correlations were lower than those in the first set

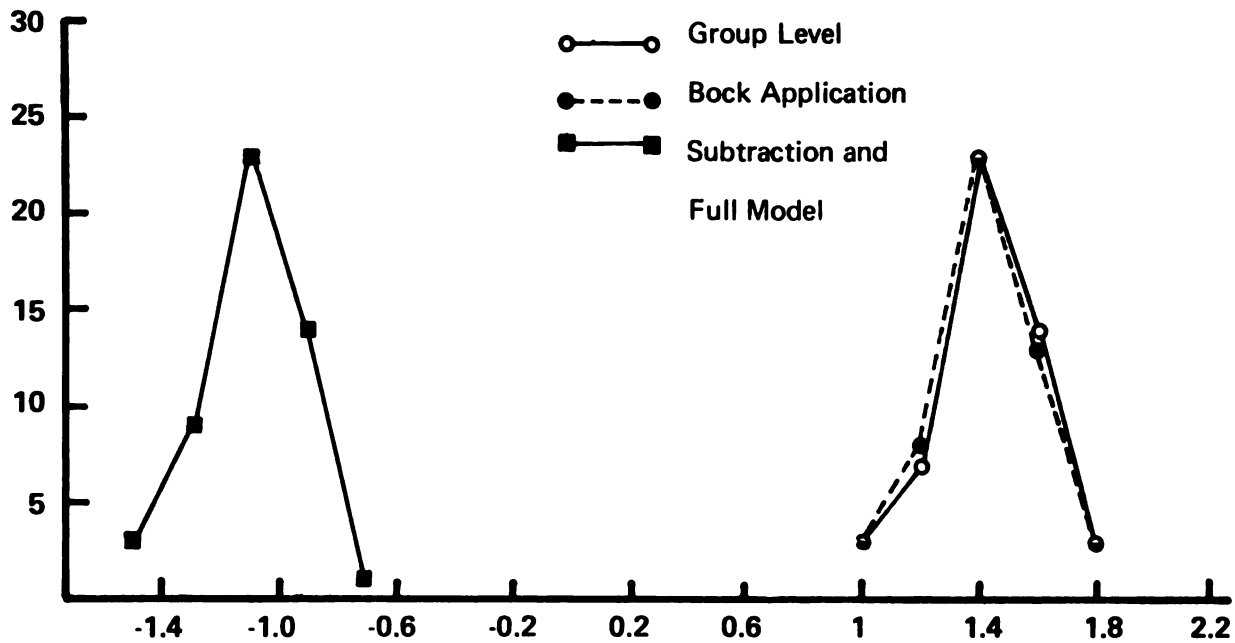

Figure 5-11 Sampling Distribution of β_{a1} From Population II-C

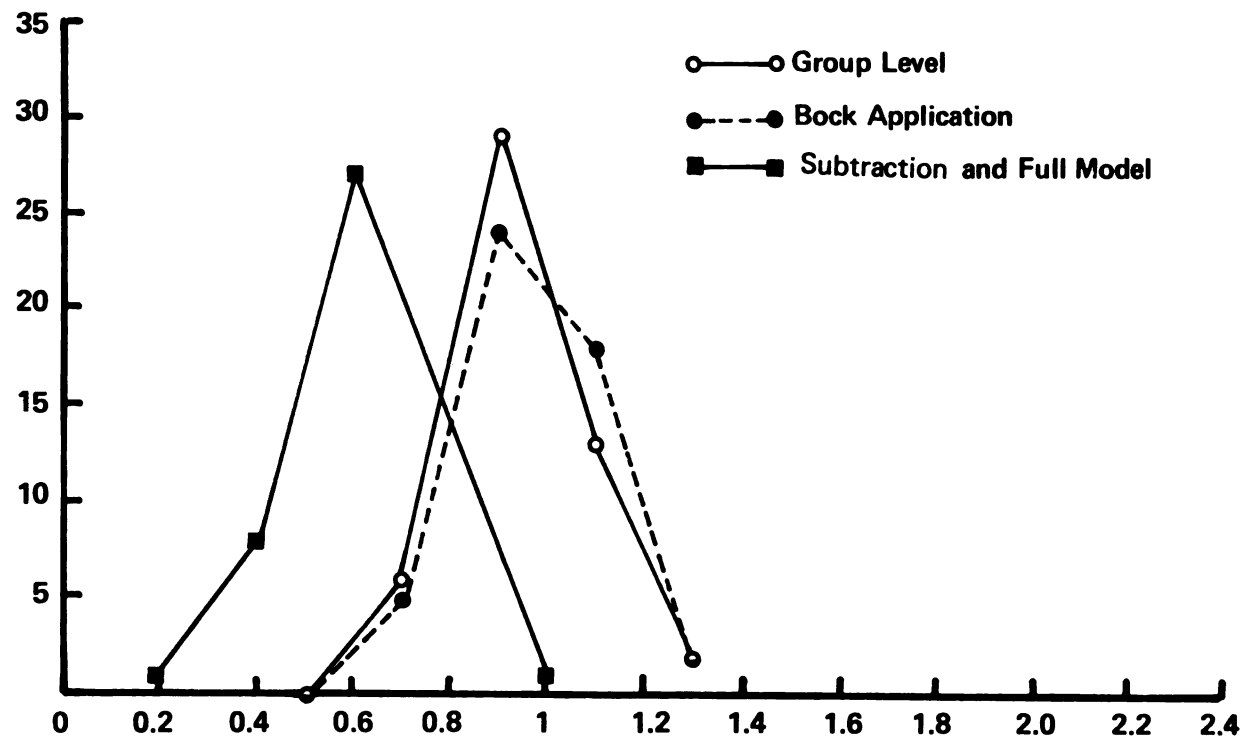

Figure 5-12 Sampling Distribution of β_{a2} From Population II-C

of data. All analysis approaches were used to analyze the second set of data in the same fashion as for the first set of data. The intra-class correlations of the second set of data are 0.30.

The ten samples for population I-C were analyzed by the two stage analysis approach, the group level analysis approach, the full model analysis approach and the subtraction approach. The results of the data analysis of the second set of data in population I-C are shown in Table 5-7. The parameter values of β_1 and β_2 were 2.53 and 0.32. The means of the estimates of β_1 and β_2 were 2.522 and 0.329 for all four approaches. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ for all four approaches. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ for all four approaches were quite small (about 0.013 and 0.010). The bias ratios of $\hat{\beta}_1$ and $\hat{\beta}_2$ for all four approaches were 0.039 and 0.081. The results of testing the hypothesis that the means of the estimates of the within regression coefficients of all 10 samples were equal to the parameters β_1 and β_2 were not significant at 0.01 level of significance ($t = -0.615$, $t = 0.900$). Therefore, all four approaches gave the same good estimates of within regression coefficients.

The means of the estimates of the regression coefficients defined for the group level variables ($\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$) analyzed by the two stage analysis approach were closer to the parameter values than when using the other three approaches. However, the results of testing the hypothesis that the means of the estimates of the regression coefficients defined for the group level variables of all 10 samples were equal to the parameters β_{z1} and β_{z2} were not significant at the 0.01 level of significance for all four approaches. The bias ratios of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ analyzed by two stage analysis approach were 0.001 and 0.007, while the

Table 5-7

Simulation Results of the Second Set of Data of Population I-C

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Two stage	2.522	0.042	0.013	-0.615	0.039
$\beta_1 = 2.53$	Group level	2.522	0.042	0.013	-0.615	0.039
	Full model	2.522	0.042	0.013	-0.615	0.039
	Subtraction	2.522	0.042	0.013	-0.615	0.039
	Two stage	0.329	0.032	0.010	0.900	0.081
$\beta_2 = 0.32$	Group level	0.329	0.032	0.010	0.900	0.081
	Full model	0.329	0.032	0.010	0.900	0.081
	Subtraction	0.329	0.032	0.010	0.900	0.081
	Two stage	4.221	7.162	2.265	0.062	0.001
$\beta_{z1} = 4.08$	Group level	3.004	6.652	2.103	-0.512	0.028
	Full model	3.004	6.652	2.103	-0.512	0.028
	Subtraction	3.004	6.652	2.103	-0.512	0.028
	Two stage	5.450	42.965	13.587	0.243	0.007
$\beta_{z2} = 2.15$	Group level	7.509	44.182	13.971	0.384	0.016
	Full model	7.509	44.182	13.971	0.384	0.016
	Subtraction	7.509	44.182	13.971	0.384	0.016

**Ratio of the estimate of the bias to mean square error.

bias ratios of $\hat{\beta}_{z1}$ and $\hat{\beta}_{z2}$ analyzed by the other three approaches were 0.028 and 0.016. We can conclude that in the situation where the group level effects were not equal to the individual level effects ($\underline{\beta}_a \neq \underline{\beta} \neq \underline{0}$) and the intraclass correlations were not high (about 0.30) all four approaches gave the correct estimates of the regression coefficients and with small bias ratios but that the two stage approach resulted in the smallest bias. The standard errors for $\hat{\underline{\beta}}_{z1}$ and $\hat{\beta}_{z2}$ for all four approaches, however, were quite high (see Table 5-7).

Twenty five samples of the second set for population II-C were analyzed by the group level analysis approach, the Bock application analysis approach, the full model analysis approach and the subtraction analysis approach. The results of the data analyses of the second set of data for population II-C are shown in Table 5-8. The parameter values of β_1 and β_2 were 0.08 and 0.76. The means of the estimates of β_1 and β_2 were 0.083 and 0.758 for all four approaches. The standard errors of $\hat{\beta}_1$ and $\hat{\beta}_2$ of all four approaches were 0.004 and 0.003. The bias ratios of $\hat{\beta}_1$ and $\hat{\beta}_2$ of all four approaches were 0.025 and 0.014. The results of testing the hypothesis that the means of the estimates of pooled within regression coefficients over the 25 samples were equal to the parameters β_1 and β_2 were not significant at 0.01 level of significance ($t = 0.025$, $t = 0.014$). Therefore, all four approaches gave the same good estimates of the pooled within regression coefficients.

The means of the estimates of the between-group regression coefficients for the individual level variables analyzed by the group level analysis approach and the Bock application analysis approach were quite similar and close to the parameter values ($\beta_{a1} = 0.05$, $\beta_{a2} = 0.95$).

Table 5-8

Simulation Results of the Second Set of Data of Population II-C

Parameters	Analysis Approach	Estimators			t	Ratio**
		Mean	SD	SE		
	Group level	0.083	0.019	0.004	0.750	0.025
$\beta_1 = 0.08$	Bock application	0.083	0.019	0.004	0.750	0.025
	Full model	0.083	0.019	0.004	0.750	0.025
	Subtraction	0.083	0.019	0.004	0.750	0.025
	Group level	0.758	0.017	0.003	-0.667	0.014
$\beta_2 = 0.76$	Bock application	0.758	0.017	0.003	-0.667	0.014
	Full model	0.758	0.017	0.003	-0.667	0.014
	Subtraction	0.758	0.017	0.003	-0.667	0.014
	Group level	0.051	0.104	0.021	0.048	0.000
$\beta_{a1} = 0.05$	Bock application	0.047	0.114	0.023	-0.130	0.001
	Full model	-0.033	0.102	0.020	-4.150*	0.408
	Subtraction	-0.033	0.102	0.020	-4.150*	0.408
	Group level	0.933	0.078	0.016	-1.063	0.047
$\beta_{a2} = 0.95$	Bock application	0.946	0.085	0.017	-0.235	0.002
	Full model	0.175	0.076	0.015	-51.667*	0.991
	Subtraction	0.175	0.076	0.015	-51.667*	0.991

*Significant at 0.01 level of significance.

**Ratio of the estimate of the bias to mean square error.

The means of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ analyzed by the group level analysis approach were 0.051 and 0.933 and when analyzed by the Bock application analysis approach they were 0.047 and 0.946. The full model analysis approach and the subtraction analysis approach gave the same average estimates of β_{a1} and β_{a2} and they were not close to the parameter values ($\bar{\hat{\beta}}_{a1} = -0.033$, $\bar{\hat{\beta}}_{a2} = 0.175$). The results of testing the hypothesis that the means of the estimates of the between-group regression coefficients were equal to the parameter values were not significant for the group level analysis approach or for the Bock application analysis approach, but they were significant ($p < 0.01$) for the full model and the subtraction analysis approach. The bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ for the group level and the Bock application analysis approaches were quite small (bias ratios of $\hat{\beta}_{a1}$ when using the group level analysis approach and the Bock application analysis approach were 0.000 and 0.001, and bias ratios of $\hat{\beta}_{a2}$ when using the group level analysis approach and the Bock application analysis approach were 0.047 and 0.002) while the bias ratio of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ analyzed by the full model and the subtraction analysis approaches were quite large (bias ratios of $\hat{\beta}_{a1}$ and $\hat{\beta}_{a2}$ with these two apporaches were 0.408 and 0.991). From this we can conclude that in the situation where the between-group regression coefficients were not equal to the within-group regression coefficients ($\underline{\beta}_a \neq \underline{\beta} \neq \underline{0}$) and the intraclass correlations were not high (0.30) the group level and Bock application analysis approaches still gave the correct estimates of the between-group regression coefficients and the full model and the subtraction analysis approaches continued to yield incorrect estimates.

CHAPTER VI

CONCLUSIONS AND RECOMMENDATIONS

The main purpose of the present study was to investigate various alternatives used to analyze hierarchical data by applying them to a set of simulated data. We determined which approach would give the best estimates of the between and within regression coefficients in terms of accuracy (the least amount of bias) and in terms of precision for various situations. The bias ratio of each estimator was computed to facilitate comparisons.

Two situations were investigated in this dissertation. The first situation was one in which there were both individual-level predictors which can be aggregated to the group-level and predictors defined only at the group level. The second situation was one in which there were only individual-level predictors which can be aggregated. For each situation, data were generated from three different populations: the first in which there were no group level effects; the second in which group level effects were equal to individual level effects; the third in which group level effects were not equal to the individual level effects.

The original intention of this study was to do a simulation study to contrast the various analysis methods. However, the simulation results suggested several patterns that imply certain relationships between the alternative approaches. Therefore, an analytical study of the relationship among the alternative approaches was used as a follow up. The results of the analytical work are presented in this section

supporting the simulation results.

The simulation results showed that all analysis approaches gave the same estimates of the within-group regression coefficients for all six cases with good precision and small bias ratios. All approaches gave the same estimates of the within-group regression coefficients because they used the same basic formula (formula 4-4) to compute the pooled within-group regression coefficients.

For the situation in which there were both individual level predictors which can be aggregated to the group level and predictors defined only at the group level, the two stage analysis approach, group level analysis approach, full model analysis approach and the subtraction analysis approach were used to analyze the data. Analytically, these four approaches can be grouped into two sets. One set includes only the two stage analysis approach. The other set includes the group level analysis approach, the full model analysis approach, and the subtraction analysis approach. Theoretically, these three approaches should give the same estimates of the regression coefficients defined for the group level variable ($\underline{\beta}_z$) by the two stage analysis approach and the other three approaches are defined by formulas (6-1) and (6-2) respectively.

$$(6\text{-}1) \quad \hat{\underline{\beta}}_z = B_z^{-1}B_{zy} - B_z^{-1}B_{zx}\,(\hat{\lambda}\hat{\underline{\beta}})$$

$$(6\text{-}2) \quad \hat{\underline{\beta}}_z = B_z^{-1}B_{zy} - B_z^{-1}B_{zx}\hat{\underline{\beta}}_a$$

where B_z, B_{zy}, and B_{zx} are the between-group sum of squares and cross product matrices of $\underline{Z}$;, $\underline{Z}$ and Y, and $\underline{Z}$ and $\underline{X}$ variables.

The simulation results showed that for all three cases, the group level analysis approach, the full model approach and the subtraction

approach gave the same estimates of $\underline{\beta}_z$ and were consistent with the theoretical results suggested above.

In the case where there were no group level effects, the two stage analysis approach gave estimates of $\underline{\beta}_z$ better than those derived from the other three approaches. Where the between-group regression coefficients were equal to the pooled within group regression coefficients, all four approaches gave essentially the same estimates of $\underline{\beta}_z$, all with comparable bias ratios. In the case where the between-group regression coefficients were neither equal to the pooled within-group regression coefficients nor to zero, the simulation results were different depending upon the value of the intraclass correlation coefficient. For the case where the intraclass coefficient was high, the two stage analysis did not give as good estimates of $\underline{\beta}_z$ as the other three approaches. However, when the intraclass correlations were more moderate in value (around 0.30) all four approaches gave the same estimates of $\underline{\beta}_z$, although the two stage approach yielded better bias ratios indicating less bias relative to mean square error.

When the situation was such that there were only individual level predictors which could be aggregated to the group level, the group level analysis approach, Bock application analysis approach, full model analysis approach and subtraction analysis approach were used to analyze the data. Theoretically, these four approaches can be grouped into three sets: first, the group level analysis approach by itself; second, the Bock application analysis approach by itself; and third, the full model analysis approach and the subtraction analysis approach. In theory, the estimates of the between-group regression coefficients ($\underline{\beta}_a$) by the full model analysis approach are equal to the differences

between the between-group regression coefficients that are estimated from the between-group sum of squares and cross products matrix and the pooled within group regression coefficients. Therefore, the estimates of $\underline{\beta}_a$ that obtain from the full model analysis approach and the subtraction analysis approach should be the same. Analytically, the relationship between the between-group regression coefficients estimated by the group level analysis approach, Bock application analysis approach are shown in equation (6-3).

$$(6\text{-}3) \quad \hat{\underline{\beta}}_a^B = \hat{\underline{\beta}}_a^G + (B^{-1} A - I)^{-1} \hat{\underline{\beta}}_a^F$$

where $\underline{\beta}_a^B$, $\underline{\beta}_a^G$ and $\underline{\beta}_a^F$ are the between-group regression coefficients estimated by the Bock application analysis approach, the group level analysis approach and the full model analysis approach, respectively, B is the within-group mean of square divided by the number of subjects in each group, A is the between-group mean of square divided by the number of subjects in each group, and I is the identity matrix. The derivation of this relationship is shown in Appendix B.

The simulation results showed that for all three cases, the full model analysis approach and subtraction analysis approach gave exactly the same estimates of the between-group regression coefficients. They were also equal to the difference between the regression coefficients that were estimated from the between-group sum of squares and cross products matrix and the pooled within-group regression coefficients which is consistent with the theoretical results. However, the estimates of $\underline{\beta}_a$ from these two approaches were not close to the parameter values. The bias ratios for the estimates resulting from these two approaches were very high. From this, we can conclude that the

subtraction and the full model approach gave totally wrong estimates of $\underline{\beta}_a$. For all three cases the group level analysis approach and the Bock application analysis approach gave good estimates of $\underline{\beta}_a$. The bias ratios for these two approaches were quite small when compared to the bias ratios for the other two approaches. When the between-group regression coefficients were equal to zero, the Bock application analysis approach gave better estimates of $\underline{\beta}_a$ than the group level analysis approach. However, when the between-group and within-group regression coefficients were equal, both approaches gave the same estimates of $\underline{\beta}_a$. For the situation where the between-group regression coefficients were not equal to the pooled within-group regression coefficients, the group level analysis approach gave better estimates of $\underline{\beta}_a$ than the Bock application analysis approach when the intraclass correlations were high (about 0.90), but the Bock application analysis approach gave the better estimates of $\underline{\beta}_a$ when the intraclass correlations were low (about 0.30).

From the simulation results, we can summarize which approach gave good estimates of the parameters for the different populations. This is shown in Table 6-1.

Table 6-1 shows the quality of the estimates of the between regression coefficients defined for the group level variables ($\underline{\beta}_z$) and the between-group regression coefficients ($\underline{\beta}_a$) under the alternative approaches in terms of accuracy (bias ratios less than 0.15). In the situation where there were both individual level predictors which were aggregated to the group level and predictors which were defined only at the group level, the two stage analysis approach gave good estimates of the regression coefficients defined for the group

Table 6-1

Quality of the Estimates for Different Populations

Populations		Analysis Approach				
		Two Stage	Group Level	Full Model	Subtraction	Bock Application
Both individual level and school level variables	No group level effects	Good*	Bad**	Bad	Bad	Not applicable
	Group level effects were equal to the individual level effects	Good	Good	Good	Good	Not applicable
	Group level effects were not equal to the individual level effects with moderate intraclass correlations	Good	Good	Good	Good	Not applicable
	Group level effects were not equal to the individual level effects with high intraclass correlations	Bad	Good	Good	Good	Not applicable

Table 6-1 (cont'd.)

Populations		Analysis Approach				
		Two Stage	Group Level	Full Model	Subtraction	Bock Application
Individual level variables only	No group level effects	Not applicable	Good	Bad	Bad	Good
	Group level effects were equal to the individual level effects	Not applicable	Good	Bad	Bad	Good
	Group level effects were not equal to the individual level effects with moderate intraclass correlations	Not applicable	Good	Bad	Bad	Good
	Group level effects were not equal to the individual level effects with high intraclass correlations	Not applicable	Good	Bad	Bad	Good

*Good is defined where the bias ratio is less than 0.15.

**Bad is defined where the bias ratio is greater than or equal to 0.15.

level variables: 1) in the case where there were no group level effects; 2) where the group level effects were equal to the individual level effects; and 3) where the group level effects were not equal to the individual level effects and when the intraclass correlations were low (about 0.30). The group level analysis, the full model and the subtraction approach gave good estimates of the regression coefficients defined for the group level variables in the case where: 1) the group level effects were not equal to the individual level effects; and 2) the group level effects were not equal to the individual level effects and when the intraclass correlations were either low (about 0.30) or high (about 0.90).

When the situation was such that there were only individual level predictors which could be aggregated to the group level, the group level analysis and Bock application approaches gave good estimates of the between-group regression coefficients for all cases. The full model and the subtraction approach gave bad estimates of the between-group regression coefficients for all cases.

In the present study, we only dealt with the simulation of specific parameter values and specific situations. We did not investigate all types of parameter values or all types of situations. Therefore, the results of this study can be generalized only to similar situations and similar parameter values.

Recommendations for Further Study

The present study deals with situations where homogeneity of within-group regression coefficients is assumed; therefore, one possible extension of the present work is an investigation of the methods of

analyzing hierarchical data which allow for heterogeneity of the within-group regression coefficients. The results of this study suggest that the intraclass correlations have an effect on estimating the between-group regression coefficients ($\underline{\beta}_a$) and the between-group regression coefficients defined for the group variables ($\underline{\beta}_z$). This would suggest the investigation of all analysis approaches that are used to analyze hierarchical data for different sets of data that are generated from populations which are described by intraclass correlations of different magnitudes. The present study, although not designed to examine this issue, and upon finding the apparent relationship, was able to suggest in a preliminary way the need for examining this issue more thoroughly.

Another avenue of future work is to apply the analytical procedures based on the methods of analysis of covariance structures for hierarchical data devised by Schmidt (1969) and Wisenbaker and Schmidt (1979) to simulated data of the sort considered in this study.

APPENDICES

APPENDIX A

COMPUTER PROGRAMS

```
      PROGRAM MYDATA(INPUT,OUTPUT=65,TAPE6=OUTPUT,TAPE5,TAPE1,TAPE2,TAPE
     +3)
C     GENERATION PROGRAM
C     SUBROUTINES NEED
C     GENEA
C     CHOL
C     CHANGE
C     GENDATA
C     COVAR
      DIMENSION SIGMA(15),T(5,5),Z(5,1),E(1500,5),A(15),TEMP(5,1)
      DIMENSION SIGMAA(15),TA(5,5),AI(1500,5),Y(1500,5),GTOTAL(5)
      DIMENSION GMS(5),YBAR(1500,5),SUM(50,5),W(15),S(15),MU(5),SV(15)
      DIMENSION SSW(15),SVB(15),SVW(15),GMB(5),PM(5),GTW(5),PSV(15)
      DIMENSION SHAT(15),GMEAN(5),SVBT(15)
      REAL MU
C     READ IN K,KW,N,NS,NT,NE,NEW,NSAM,SIGMA,SIGMAA,MU
C     K=NO. OF VARIABLES
C     KW=NO. OF VARIABLES FOR WITHIN SCHOOL
C     N=NO. OF SUBJECTS WITHIN EACH SCHOOL
C     NS=NO. OF SCHOOL
C     NT=NO. OF TOTAL SUBJECTS
C     NE=NO. OF ELEMENTS IN COVARIANCE MATRIX
C     NEW=NO. OF ELEMENTS IN WITHIN COVARIANCE MATRIX
C     NSAM=NO. OF SAMPLES
      READ(5,10)K,KW,N,NS,NT,NE,NEW,NSAM
   10 FORMAT(8I5)
      READ(5,15)(SIGMA(I),I=1,NEW),(SIGMAA(I),I=1,NE),(MU(I),I=1,K)
   15 FORMAT(6F10.4,/,8F10.4,/,7F10.4,/,5F10.4)
C     WRITE K,KW,N,NS,NT,NE,NEW,NSAM,SIGMA,SIGMAA,MU
      WRITE(6,20)K,KW,N,NS,NT,NE,NEW,NSAM,(SIGMA(I),I=1,NEW),
     +(SIGMAA(I),I=1,NE),(MU(I),I=1,K)
   20 FORMAT(*1DATA INFORMATION*,//,5X,*NO. OF VARIABLES = *,I5,/,5X,*NO
     +. OF WITHIN SCHOOL VARIABLES = *,I5,/,5X,*NO. OF SUBJECTS WITHIN S
     +CHOOL = *,I5,/,5X,*NO. OF SCHOOLS = *,I5,/,5X,*NO. OF TOTAL SUBJEC
     +TS = *,I5,/,5X,*NO. OF ELEMENTS IN COVARIANCE MATRIX = *,I5,/,5X,*
     +NO. OF ELEMENTS IN WITHIN COVARIANCE MATRIX = *,I5,/,5X,*NO. OF SA
     +MPLES = *,I5,/,*0THE WITHIN COVARIANCE MATRIX*,//,5X,F10.4,/,5X,2F
     +10.4,/,5X,3F10.4,/,*0THE BETWEEN COVARIANCE MATRIX*,//,5X,F10.4,/,
     +5X,2F10.4,/,5X,3F10.4,/,5X,4F10.4,/,5X,5F10.4,/,*0POPULATION MEAN
     +=*,5F10.4)
C     START GENERATING DATA
      DO 100 IJK=1,NSAM
C     PRINT SAMPLE NUMBER
      WRITE(6,21)IJK
   21 FORMAT(*0SAMPLE NO. *,I2)
C     GENERATE E(I,J)
      CALL GENEA(KW,NT,SIGMA,E,T)
C     WRITE TWO MORE COLUMNS FOR WITHIN COVARIANCE
      DO 750 I=1,NT
        E(I,4)=0.
        E(I,5)=0.
  750 CONTINUE
C     WRITE CHOLESKY FAACTOR OF WITHIN COVARIANCE
      WRITE(6,25)
   25 FORMAT(*0THE CHOLESLY FACTOR OF WITHIN COVARIANCE*)
      DO 110 I=1,KW
        WRITE(6,30)(T(I,J),J=1,KW)
   30   FORMAT(*0*,3F10.4)
  110 CONTINUE
C     GENERATE AI(I)
      CALL GENEA(K,NS,SIGMAA,AI,TA)
C     WRITE CHOLESKY FACTOR OF BETWEEN COVARIANCE
      WRITE(6,35)
   35 FORMAT(*0THE CHOLESKY FACTOR OF BETWEEN COVARIANCE*)
      DO 120 I=1,K
```

```
      WRITE(6,40)(TA(I,J),J=1,K)
   40   FORMAT(*0*,5F10.4)
  120 CONTINUE
C     GENERATE Y(I,J)
      CALL GENDATA(K,N,NS,NE,NT,MU,E,AI,Y,YBAR,PM,PSV,GMB,SVB,GMEAN,SV,S
     +HAT)
C     PRINT POOLED MEAN OF EACH VARIABLE
      WRITE(6,45)(PM(J),J=1,K)
   45 FORMAT(*0VECTOR OF POOLED MEAN =*,5(F10.4,3X))
C     PRINT SAMPLE POOLED WITHIN COVARIANCE
      WRITE(6,50)(PSV(L),L=1,NE)
   50 FORMAT(*0SAMPLE POOLED WITHIN COVARIANCE MATRIX*,//,5X,F10.4,/,5X,
     +2(F10.4,3X),/,5X,3(F10.4,3X),/,5X,4(F10.4,3X),/,5X,5(F10.4,3X))
C     PRINT GRAND MEAN OF EACH VARIABLE,SCHOOL MEAN IS UNIT OF ANALYSIS
      WRITE(6,60)(SVB(L),L=1,NE)
   60 FORMAT(*0SAMPLE BETWEEN COVARIANCE MATRIX*,//,5X,F10.4,/,5X,2(F10.
     +4,3X),/,5X,3(F10.4,3X),/,5X,4(F10.4,3X),/,5X,5(F10.4,3X))
C     PRINT GRAND MEAN OF EACH VARIAABLE
      WRITE(6,65)(GMEAN(J),J=1,K)
   65 FORMAT(*0VECTOR OF GRAND MEAN =*,5(F10.4,3X))
C     PRINT SAMPLE COVARIANCE MATRIX
      WRITE(6,70)(SV(L),L=1,NE)
   70 FORMAT(*0SAMPLE COVARIANCE MATRIX*,//,5X,F10.4,/,5X,2(F10.4,3X),/,
     +5X,3(F10.4,3X),/,5X,4(F10.4,3X),/,5X,4(F10.4,3X))
C     PRINT PURE BETWEEN COVARIANCE
      WRITE(6,75)(SHAT(L),L=1,NE)
   75 FORMAT(*0SAMPLE PURE BETWEEN COVARIANCE*,//,5X,F10.4,/,5X,2(F10.4,
     +3X),/,5X,3(F10.4,3X),/,5X,4(F10.4,3X),/,5X,5(F10.4,3X))
C     PRINT Y AND YBAR ON TAPE1
      II=1
      L=N
      DO 125 M=1,NS
        DO 130 I=II,L
          WRITE(1,80)M,I,(Y(I,J),J=1,K),(YBAR(M,J),J=2,3)
   80     FORMAT(I2,1X,I4,7(F10.4))
  130   CONTINUE
        II=II+N
        L=L+N
  125 CONTINUE
      ENDFILE 1
C     PRINT SCHOOL MEAN ON TAPE2
      DO 135 M=1,NS
        WRITE(2,85)M,(YBAR(M,J),J=1,K)
   85   FORMAT(I2,5(3X,F10.4))
  135 CONTINUE
      ENDFILE 2
C     PRINT POOLED WITHIN COVARIANCE ON TAPE3
      WRITE(3,90)PSV(1),PSV(2),PSV(4),PSV(2),PSV(3),PSV(5),PSV(4),PSV(5)
     +,PSV(6)
   90 FORMAT(3F10.4,/,3F10.4,/,3F10.4)
      ENDFILE 3
  100 CONTINUE
      END
      SUBROUTINE GENEA(K,N,SIGMA,Y,T)
C     GENERATION PROGRAM FOR A(I) AND E(I,J)
C     READ IN SIGMA
C     FIND CHOLESKY FACTOR OF SIGMA = T
C     GENERATE 5 VARIABLES DISTRIBUTED N(0,1) - Z-5X1
C     TRANSFORM A(I)=TZ, A(I) DISTRIBUTED N(0,SIGMA-A)
C
C     SUBROUTINES NEED
C
```

```
C       CHOL
C       CHANGE
C
        DIMENSION SIGMA(15),T(5,5),Z(5,1),Y(1500,5),A(15),TEMP(5,1)
C       SET GENERATOR PARAMETER
        A1=3.949846138
        A3=0.252408784
        A5=0.076542912
        A7=0.008355968
        A9=0.029899776
        CALL RANSET(CLOCK(DUMMY))
C       FIND CHOLESKY FACTOR OF SIGMA =T AND DETERMINANT OF T
        CALL CHOL(SIGMA,A,K,DET)
        CALL CHANGE(A,T,K)
        DO 700 II=1,N
C       GENERATE 12 RANDOM NUMBER DISTRIBUTED U(0,1)
          DO 200 J=1,K
            B=0.
            DO 300 IN=1,12
              RX=RANF(DUMMY)
              B=B+RX
  300       CONTINUE
C       TRANSFORM UNIFORM RANDOM NUMBER TO Z VARIABLE DISTRIBUTED N(0,1)
            R=(B-6.)/4.
            RS=R*R
            Z(J,1)=((((A9*RS+A7)*RS+A5)*RS+A3)*RS+A1)*R
  200     CONTINUE
C       TRANSFORM Z VARIABLE TO Y VARIABLE DISTRIBUTED N(0,SIGMA)
          DO 400 JJ=1,K
            X=0.
            DO 500 KK=1,K
              X=X+T(JJ,KK)*Z(KK,1)
  500       CONTINUE
            TEMP(JJ,1)=X
  400     CONTINUE
          DO 600 J=1,K
            Y(II,J)=TEMP(J,1)
  600     CONTINUE
  700   CONTINUE
        RETURN
        END
        SUBROUTINE CHOL(SIGMA,A,K,DET)
C
C       SIGMA AN K BY K SYMMETRIC MATRIX
C       A       AN ARRAY OF AT LEAST K*(K+1)/2 LOCATIONS
C       K       NUMBER OF ROWS IN SIGMA
C       DET     THE DETERMINANT OF A
        DIMENSION SIGMA(15),A(15)
        X=SQRT(SIGMA(1))
        DET=X
        A(1)=X
        K1=K-1
        KC=1
        IFIR=1
        DO 101 J=1,K1
          KC=KC+J
          A(KC)=SIGMA(KC)/X
  101   CONTINUE
        DO 105 I=1,K1
          IFIR=IFIR+I
          KC=IFIR
          X=0.
```

```
      DO 102 J=1,I
        X=X+A(KC)**2
        KC=KC+1
  102 CONTINUE
      X=SQRT(SIGMA(KC)-X)
      DET =DET*X
      A(KC)=X
      II=I+1
      IF(II.EQ.K)RETURN
      JC=IFIR
      DO 103 J=II,K1
        JC=JC+J
        IC=JC
        KC=IFIR
        Y=0.
        DO 104 L=1,I
          Y=Y+A(IC)*A(KC)
          KC=KC+1
          IC=IC+1
  104   CONTINUE
        A(IC)=(SIGMA(IC)-Y)/X
  103 CONTINUE
  105 CONTINUE
      RETURN
      END
      SUBROUTINE CHANGE(A,T,K)
C
C     A MATRIX TO BE CONVERTED
C     T ARRAY WHERE CONVERTED MATRIX WILL BE STOKED
C     K DIMENSION OF MATRIX
C
C     CHANGE TO SQUARE MATRIX
C
      DIMENSION A(15),T(5,5)
      L=K+1
      LL=(L*K)/2+1
      DO 41 J=1,K
        JR=L-J
        DO 42 I=1,JR
          IR=JR-I+1
          LL=LL-1
          T(JR,IR)=A(LL)
   42   CONTINUE
   41 CONTINUE
      DO 43 J=2,K
        L=J-1
        DO 44 I=1,L
          T(I,J)=0.
   44   CONTINUE
   43 CONTINUE
      RETURN
      END
      SUBROUTINE GENDATA(K,N,NS,NE,NT,MU,E,A,Y,YBAR,PM,PSV,GMB,SVB,GMEAN
     +,SV,SHAT)
C     GENERATION PROGRAM FOR Y(I,J)
C     Y(I,J)=MU+A(I)+E(I)
C     THARE ARE 5 VARIABLES FOR EACH SUBJECTS
C     VECTOR OF Y DISTRIBUTED N(MU,SIGMA)
      DIMENSION E(1500,5),A(1500,5),Y(1500,5),GTOTAL(5),GMEAN(5),GMS(5)
      DIMENSION YBAR(1500,5),SUM(50,5),W(15),S(15),MU(5),SV(15),SSW(15)
      DIMENSION SVB(15),SVW(15),GMB(5),PM(5),GTW(5),PSV(15),SHAT(15)
      DIMENSION SVRT(15)
      REAL MU
```

```
C     K= NO. OF VARIABLES, N=NO. OF SUBJECTS IN EACH SCHOOL
C     NS=NO. OF SCHOOLS, NE=NO. OF ELEMENTS, NT=TOTAL NUMBER OF SUBJECTS
C     COMPUTE Y(I,J)
        II=1
        L=N
      DO 100 M=1,NS
        DO 200 I=II,L
          DO 250 J=1,K
            Y(I,J)=MU(J)+A(M,J)+E(I,J)
  250     CONTINUE
  200   CONTINUE
        II=II+N
        L=L+N
  100 CONTINUE
C     COMPUTE SUM AND MEAN FOR EACH SCHOOL
      II=1
      L=N
      AN=N
      DO 450 M=1,NS
        DO 400 J=1,K
          SUM(M,J)=0.
          DO 500 KK=II,L
            SUM(M,J)=SUM(M,J)+Y(KK,J)
  500     CONTINUE
          YBAR(M,J)=SUM(M,J)/AN
  400   CONTINUE
        II=II+N
        L=L+N
  450 CONTINUE
C     COMPUTE POOLED MEAN
      ANT=NT
      DO 94 J=1,K
        GTW(J)=0.0
        DO 93 M=1,NS
          GTW(J)=GTW(J)+SUM(M,J)
   93   CONTINUE
        PM(J)=GTW(J)/ANT
   94 CONTINUE
C     COMPUTE POOLED WITHIN COVARIANCE MATRIX
      AN1=N-1
      DO 97 J=1,NE
        SSW(J)=0.0
   97 CONTINUE
      II=1
      L=N
      DO 98 I=1,NS
        CALL COVAR(II,L,Y,K,N,NE,SVW,GMS)
        DO 99 J=1,NE
          SSW(J)=SSW(J)+(AN1*SVW(J))
   99 CONTINUE
        II=II+N
        L=L+N
   98 CONTINUE
      PN=NT-NS
      DO 96 J=1,NE
        PSV(J)=SSW(J)/PN
   96 CONTINUE
C     COMPUTE SAMPLE BETWEEN COVARIANCE
      II=1
      L=NS
      CALL COVAR(II,L,YBAR,K,NS,NE,SVBT,GMB)
      AN=N
```

```
      DO 451 I=1,NE
       SVB(I)=SVBT(I)*AN
  451 CONTINUE
C     COMPUTE SAMPLE TOTAL COVARIANCE MATRIX
      II=1
      L=NT
      CALL COVAR(II,L,Y,K,NT,NE,SV,GMEAN)
C     COMPUTE THE ESTIMATION OF PURE BETWEEN COVARIANCE
      AN=N
      DO 130 I=1,NE
        SHAT(I)=(SVB(I)-PSV(I))/AN
  130 CONTINUE
      RETURN
      END
      SUBROUTINE COVAR(II,L,Y,K,NT,NE,SV,GMEAN)

C     COMPUTE MEAN AND COVARIANCE MATRIX
      DIMENSION Y(1500,5),SV(15),GMEAN(5),GTOTAL(5),W(15),S(15)
C     COMPUTE SUM AND MEAN
      ANT=NT
      DO 600 J=1,K
        GTOTAL(J)=0.
        DO 700 I=II,L
          GTOTAL(J)=GTOTAL(J)+Y(I,J)
  700   CONTINUE
        GMEAN(J)=GTOTAL(J)/ANT
  600 CONTINUE
C     COMPUTE YBAR BY YBAR TRANSPOSE
      IC=0
      DO 750 J=1,K
        DO 710 M=1,J
          IC=IC+1
          W(IC)=GMEAN(J)*GMEAN(M)
  710   CONTINUE
  750 CONTINUE
C     COMPUTE NT BY (YBAR BY YBAR TRANSPOSE)
      ANT=NT
      DO 800 I=1,NE
        W(I)=W(I)*ANT
  800 CONTINUE
C     COMPUTE Y TRANSPOSE Y
      IC=0
      DO 925 J=1,K
        DO 900 M=1,J
          X=0.
          DO 950 KK=II,L
            X=X+Y(KK,J)*Y(KK,M)
  950     CONTINUE
          IC=IC+1
          S(IC)=X
  900   CONTINUE
  925 CONTINUE
C     COMPUTE SAMPLE VARIANCE COVARIANCEMATRIX
      ANT1=NT-1
      DO 955 J=1,NE
        S(J)=S(J)-W(J)
        SV(J)=S(J)/ANT1
  955 CONTINUE
      RETURN
      END
```

```
DISCRIMINATION ANALYSIS (  FINN MANOVA PROGRAM                    TO
GET CHARACTERISTIC ROOT AND VECTOR TO USE IN BOCK APPROACH
   3    1    1    1                        20                          1    1
SCHOOL  50
SAMPLE NUMBER 2
FINISH
(I2,5X,3(F10.4))
Y      X1     X2
  50
C0,
C1,
C2,
C3,
C4,
C5,
C6,
C7,
C8,
C9,
C10,
C11,
C12,
C13,
C14,
C15,
C16,
C17,
C18,
C19,
C20,
C21,
C22,
C23,
C24,
C25,
C26,
C27,
C28,
C29,
C30,
C31,
C32,
C33,
C34,
C35,
C36,
C37,
C38,
C39,
C40,
C41,
C42,
C43,
C44,
C45,
C46,
C47,
C48,
C49,
   3                 1    1
-1,49.
                                                                    STOP
```

```
      PROGRAM BOCK(INPUT,OUTPUT=65,TAPE5,TAPE6=OUTPUT)

C     SAMPLE NUMBER 50

C     ESTIMATE POSITIVE SEMI DEFINITE BETWEEN COVARIANCE
C     USE IMSL SUBROUTINE
      DIMENSION WKAREA(200),TINVT(3,3),TINV(3,3),AI(3,3),T(3,3)
      DIMENSION PHI(3,3),PHIMI(3,3),RES(3,3),SIGMA(3,3),RES1(3,3)
      DIMENSION SXX(2,2),SXY(2,1),SXXINV(2,2),BHAT(2,1)
      DIMENSION SYX(1,2),SE(2)
C     N=NO. OF SUBJECTS IN EACH SCHOOL
C     K=NO. OF DIMENSION OF MATRICES
      N=30
      K=3
C     READ IN T(I,J),PHI(I,J)
      DO 100 I=1,K
        READ(5,10)(T(I,J),J=1,K),(PHI(I,J),J=1,K)
   10   FORMAT(6F10.6)
C     CREATE IDENTITY MATRIX
        DO 200 J=1,K
        IF (I.EQ.J) THEN
          AI(I,J)=1.
        ELSE
          AI(I,J)=0.
        ENDIF
  200   CONTINUE
  100 CONTINUE
C     PRINT T(I,J),PHI(I,J),AI(I,J)
      WRITE(6,20)
   20 FORMAT(*1MATRIX OF CHARACTERISTIC VECTOR*)
      DO 300 I=1,K
        WRITE(6,25)(T(I,J),J=1,K)
   25   FORMAT(*0*,(3X,3(F10.6,3X)))
  300 CONTINUE
      WRITE(6,30)
   30 FORMAT(*0MATRIX OF PHI*)
      DO 350 I=1,K
        WRITE(6,25)(PHI(I,J),J=1,K)
  350 CONTINUE
      WRITE(6,50)
   50 FORMAT(*0IDENTITY MATRIX*)
      DO 375 I=1,K
        WRITE(6,25)(AI(I,J),J=1,K)
  375 CONTINUE
C     COMPUTE PHI-I
      DO 400 I=1,K
        DO 500 J=1,K
          PHIMI(I,J)=PHI(I,J)-AI(I,J)
  500   CONTINUE
  400 CONTINUE
C     COMPUTE INVERSE OF T
      CALL LINV2F(T,3,3,TINV,0,WKAREA,IER)
C     CREATE TINVERSE TRANSPOSE
      DO 600 I=1,K
        DO 700 J=1,K
          TINVT(I,J)=TINV(J,I)
  700   CONTINUE
  600 CONTINUE
C     PRINT INVERSE OF T
      WRITE(6,60)
   60 FORMAT(*0INVERSE OF T*)
      DO 725 I=1,K
        WRITE(6,25)(TINV(I,J),J=1,K)
  725 CONTINUE
```

```
C     PRINT TRANSPOSE OF TINVERSE
      WRITE(6,70)
   70 FORMAT(*0TRANSPOSE OF TINVERSE*)
      DO 750 I=1,K
        WRITE(6,25)(TINVT(I,J),J=1,K)
  750 CONTINUE
C     COMPUTE THE PRODUCT OF TINVERSE TRANSPOSE AND PHI-I
      CALL VMULFF(TINVT,PHIMI,3,3,3,3,3,RES1,3,IER)
C     COMPUTE THE PRODUCT OF TINVERSE TRANSPOSE,PHI-I AND TINVERSE
      CALL VMULFF(RES1,TINV,3,3,3,3,3,RES,3,IER)
C     COMPUTE POSITIVE SEMI DEFINITEBETWEEN COVARIANCE(SIGMA-A HAT)
      AN=N
      DO 800 I=1,K
        DO 850 J=1,K
          SIGMA(I,J)=RES(I,J)/AN
  850   CONTINUE
  800 CONTINUE
C     PRINT POSITIVE SEMI DEFINITE BETWEEN COVARIANCE
      WRITE(6,80)
   80 FORMAT(*0POSITIVE SEMI DIFINITE BETWEEN COVARIANCE MATRICES(SIGMA-
     +HAT)*)
      DO 900 I=1,K
        WRITE(6,25)(SIGMA(I,J),J=1,K)
  900 CONTINUE
      SYY=SIGMA(1,1)
      SXY(1,1)=SIGMA(2,1)
      SXY(2,1)=SIGMA(3,1)
      SXX(1,1)=SIGMA(2,2)
      SXX(1,2)=SIGMA(2,3)
      SXX(2,1)=SIGMA(3,2)
      SXX(2,2)=SIGMA(3,3)
      CALL BETAHAT(SYY,SXY,SXX)
      END
      SUBROUTINE BETAHAT(SYY,SXY,SXX)
C     COMPUTE BETWEEN SLOPE BY BOCK APPLICATION APPROACH
C     USE IMSL SUBROUTINE
      DIMENSION SXX(2,2),SXY(2,1),SXXINV(2,2),BHAT(2,1),WKAREA(200)
      DIMENSION SYX(1,2),SE(2)
C     N=NO. OF SUBJECTS
C     K=NO. OF X
      N=50
      K=2
C     PRINT SXX
      WRITE(6,20)
   20 FORMAT(*1MATRIX OF SXX*)
      DO 200 I=1,K
        WRITE(6,30)(SXX(I,J),J=1,K)
   30   FORMAT(*0*,5X,2(F10.4,3X))
  200 CONTINUE
C     PRINT SXY
      WRITE(6,40)
   40 FORMAT(*0MATRIX OF SXY*)
      DO 300 I=1,K
        WRITE(6,50)(SXY(I,1))
   50   FORMAT(*0*,5X,F10.4)
  300 CONTINUE
C     COMPUTE INVERSE OF SXX
      CALL LINV2F(SXX,2,2,SXXINV,0,WKAREA,IER)
C     PRINT INVERSE OF SXX
      WRITE(6,60)
   60 FORMAT(*0INVERSE OF SXX*)
```

APPENDIX B

RELATIONSHIP OF THE BETWEEN-GROUP REGRESSION COEFFICIENTS FROM VARIOUS ANALYSIS APPROACHES

Relationship of the Between-Group Regression Coefficients From Various Analysis Approaches

To estimate the between-group regression coefficients, Bock application approach uses the estimated between-group variance-covariance matrix, Σ_a. An unbiased estimate of Σ_a is:

$$\hat{\Sigma}_a = \frac{1}{n}(S_a - S)$$

where S_a is the matrix of the between-group mean squares,

S is the matrix of the within-group mean squares,

and n is the number of subjects within each group.

Denote:

$$A = \frac{\sum_{i=1}^{I} n(\bar{X}_i^{(k)} - \bar{\bar{X}}^{(k)})(\bar{X}_i^{(k')} - \bar{\bar{X}}^{(k')})}{n(I-1)} \qquad KxK$$

$$B = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n} (X_{ij}^{(k)} - \bar{X}_i^{(k)})(X_{ij}^{(k')} - \bar{X}_i^{(k')})}{nI(n-1)} \qquad KxK$$

$$\underline{a} = \frac{\sum_{i=1}^{I} n(\bar{X}_i^{(k)} - \bar{\bar{X}}^{(k)})(Y_i - \bar{\bar{Y}})}{n(I-1)} \qquad Kx1$$

and $$\underline{b} = \frac{\sum_{i=1}^{I}\sum_{j=1}^{n} (X_{ij}^{(k)} - \bar{X}_i^{(k)})(Y_{ij} - \bar{Y}_i)}{nI(n-1)} \qquad Kx1$$

where I is the number of groups.

Then $\hat{\Sigma}_a$ can be written as

$$\hat{\Sigma}_a = \begin{bmatrix} \hat{\sigma}_a(y)^2 & \hat{\Sigma}_a(xy)^1 \\ \hat{\Sigma}_a(xy) & \hat{\Sigma}_a(x) \end{bmatrix}$$

$$= \begin{bmatrix} \hat{\sigma}_a(y)^2 & (\underline{a} - \underline{b})' \\ (\underline{a} - \underline{b}) & (A - B) \end{bmatrix}$$

where $$\hat{\sigma}_a(y)^2 = \frac{1}{n} \frac{\Sigma n(\bar{Y}_i - \bar{Y})^2}{n(I - 1)} - \frac{\Sigma\Sigma(Y_{ij} - \bar{Y}_i)^2}{n(I - 1)}$$

$$= \frac{1}{n} \text{MS(Between)} - \text{MS(Within)}$$

The least squares estimate of the between-group regression coefficient can then be written as

$$\underline{\hat{\beta}}_a^B = \hat{\Sigma}_a(x)^{-1}\hat{\Sigma}_a(xy) = (A - B)^{-1} (a - b).$$

While the matrices A and B are in general non-singular, the difference matrix (A - B) may not be non-singular. Thus, Bock (1968) proposed to use the orthogonal decomposition of (A - B) and retain only those eigen values and eigen vectors that were statistically significant to construct the "inverse" of $\hat{\Sigma}_a(x)$.

In order to relate the estimated between-group regression coefficient to those obtained from the other approaches, A - B is also assumed to be non-singular so that $(A - B)^{-1}$ exists. Furthermore, both A and B are assumed to be non-singular. The least squares estimate of the between-group regression coefficient is then:

$$\hat{\beta}_a^B = (A - B)^{-1}(a - b)$$
$$= (A - B)^{-1}a - (A - B)^{-1}b$$
$$= (A(I - A^{-1}B))^{-1}a - (B(B^{-1}A - I))^{-1}b$$
$$= (I - A^{-1}B)^{-1}A^{-1}a - (B^{-1}A - I)^{-1}B^{-1}b$$

But $A^{-1}a = \underline{\hat{\beta}}_a^G$ is the between-group regression coefficient estimated from the group level apporach, and $B^{-1}b = \underline{\hat{\beta}}$ is the pooled within group regression coefficient. Hence,

$$\underline{\hat{\beta}}_a^B = (I - A^{-1}B)^{-1}\underline{\hat{\beta}}_a^G - (B^{-1}A - I)^{-1}\,\underline{\hat{\beta}}.$$

Applying a theroem presented by Nobel (1969, Theorem 5.22, p. 147), $(I - A^{-1}B)^{-1}$ can be written as:

$$(I - A^{-1}B)^{-1} = I + (B^{-1}A - I)^{-1}.$$

Thus,

$$\underline{\hat{\beta}}_a^B = (I + (B^{-1}A - I)^{-1}\,\underline{\hat{\beta}}_a^G - (B^{-1}A - I)^{-1}\,\underline{\hat{\beta}}$$
$$= \underline{\hat{\beta}}_a^G + (B^{-1}A - I)\underline{\hat{\beta}}_a^G - (B^{-1}A - I)^{-1}\,\underline{\hat{\beta}}$$
$$= \underline{\hat{\beta}}_a^G + (B^{-1}A - I)^{-1}(\underline{\hat{\beta}}_a^G - \underline{\hat{\beta}})$$
$$\underline{\hat{\beta}}_a^B = \underline{\hat{\beta}}_a^G + (B^{-1}A - I)^{-1}\underline{\hat{\beta}}_a^F$$

where $\underline{\hat{\beta}}_a^F$ is the between-group regression coefficient obtained from the full model analysis approach, and $\underline{\hat{\beta}}_a^B$ is the between-group regression coefficient obtained from the Bock application approach.

BIBLIOGRAPHY

BIBLIOGRAPHY

Bock, R. D. and Vandenberg, S. G. Components of Heritable Variation in Mental Test Scores. Progress in Human Behavior Genetics. S. G. Vandenberg (Ed.). Baltimore: Johns Hopkins Press, 1968.

Bidwell, Charles E. and Kasarda, J. D. School District Organization and Student Achievement. American Sociological Review, 1975, 40, 55-70.

Burstein, Leigh. Assessing Differences Between Grouped and Individual-Level Regression Coefficients. Paper presented at the Annual Meeting of the American Educational Research Association, 1976.

Burstein, Leigh. Three Key Topics in Regression-Based Analysis of Multilevel Data from Quasi-Experiments and Field Studies. Paper presented at Institute for Research on Teaching, Michigan State University, 1977.

Burstein, Leigh and Miller, M. David. The Use of Within-Group Slope as Indices of Group Outcomes. Paper presented at the Annual Meeting of the American Educational Research Association, 1979.

Burstein, Leigh, Linn, Robert L. and Capell, Frank J. Analyzing Multilevel Data in the Presence of Heterogeneous Within-Class Regression. Journal of Educational Statistics, 1978, 3, 347-383.

Control Data Corporation. Fortran Extended Version 4 Reference Manual, California, Control Data Corporation, 1978.

Cronbach, L. J. and Webb, N. Between-Class and Within-Class Effects in a Reported Aptitude X Treatment Interaction: Reanalysis of a Study by G. L. Anderson. Journal of Educational Psychology, 1975, 67, 717-724.

Finn, J. D. MULTIVARIANCE: Univariate and Multivariate Analysis of Variance, Covariance and Regression. Ann Arbor, Mich.: National Educational Resources, Inc., 1972.

Hannan, Michael T., and Young, Alice A. On Certain Simularities in the Estimation of Multi-Wave Panels and Panels and Multi-Level Cross-Sections. Paper prepared for the conference on Methodology for Aggregating Data in Educational Research, 1976.

Hannan, Michael T., Freeman, John, and Meyer, John W. Specification of Models of Organizational Effectiveness: Comment on Bidwell and Kasarda. American Sociological Review, 1976, 41, 136-143.

IMSL LIBRARY. The IMSL Library Volume 3. Houston, International Mathematical & Statistical Libraries, Inc., 1979.

Keesling, J. W. Components of Variance Models in Multilevel Analysis. Paper prepared for presentation at a conference on Methodology for the National Institute of Education, 1976.

Keesling, J. W. Some Explorations in Multilevel Analysis. Santa Monica: System Development Corporation, 1977.

Keesling, J. W. and Wiley, David E. Regression Models for Hierarchical Data. Paper presented at the Annual Meeting of the Psychometric Society, 1974.

Knuth, Donald E. Semi-numerical Algorithms: The Art of Computer Programming. Mass.: Addison, Wesley Publishing Co., 1968.

Noble, Ben. Applied Linear Algebra. Englewood Cliff, N.J.: Prentice-Hall, 1969.

Rock, Donald A., Baird, Leonard L. and Linn, Robert L. Interaction Between Colleges Effects and Students' Aptitudes. American Educational Research Journal, 1972, 10, 149-161.

Scheifley, Verda M. Analysis of Repeated Measures Data: A Simulation Study. Doctoral Dissertation, Michigan State University, 1974.

Scheifley, Verda M. and Schmidt, William H. Jeremy D. Finn's Multivariance-Univariate and Multivariate Analysis of Variance, Covariance, and Regression Modified and adopted for use on the CDC 6500. Occasional Paper No. 22, Office of Research Consultation, Michigan State University, 1973.