This is to certify that the

dissertation entitled

Large and Small Sample Properties of
Maximim Likelihood Estimates for the
Hierarchical Linear Model

presented by

Dina Bassiri

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Counseling, Ed. Psychology, & Special Education

Stephen W. Raudenbush
Major professor

Date 11-4-88

LARGE AND SMALL SAMPLE PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES FOR THE HIERARCHICAL LINEAR MODEL

by

Dina Bassiri

A DISSERTATION

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

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational
Psychology and Special Education

1988

ABSTRACT

LARGE AND SMALL SAMPLE PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES FOR THE HIERARCHICAL LINEAR MODEL

by

Dina Bassiri

The multilevel character of educational data have implications of a general methodological nature. Interest in these methodological problems has recently been stimulated by the development of the E.M. algorithmic approach to variance component models. The EM algorithm produces maximum likelihood estimates for variance components with known large sample properties. That is, the estimates are consistent, asymptotically efficient with known large sample normal distributions. However, at present little is known about the small sample behavior of the parameter estimates.

The primary purpose of this Monte Carlo investigation is to understand the properties of maximum likelihood estimates in small and moderate samples using a two stage hierarchical linear model with standardized normal predictors at both levels of hierarchy (i.e., using a standardized two-stage hierarchical linear model). Specifically, this research will investigate the effects of variance estimation via the EM algorithm on properties of parameter estimates at the

second stage of the hierarchy, that is, the macro or fixed effects, γ_{00}, γ_{01}, γ_{10}, and γ_{11}. These are the regression coefficients in the equations for the mean and slope at the second stage of the hierarchy. A secondary purpose is to evaluate the robustness and power of asymptotic Z-tests of the macro parameters under various conditions determined by the number of groups, the group size, and the effect size.

The following are the major conclusions drawn from the investigation. (1) Macro estimators are unbiased, consistent, and asymptotically efficient with asymptotically known normal distribution. (2) Error estimates of macro parameters are considerably affected by the number of groups, but not so much by the group size. (3) Precision of macro parameters is directly proportional to the number of groups and inversely proportional to intraclass correlation coefficient. Increasing group size increases precision as well. Yet, the effect of one is not proportional to that of the other. (4) Micro parameter variance estimator for the slope and intercept of the first stage regression model are biased but consistent and asymptotically efficient. Increasing the number of groups has a determinative effect on the parameter variance in slopes, but parameter variance in intercepts is more influenced by group size. (5) Within group error variance estimates (σ^2) are unbiased, consistent, asymptotically efficient and are considerably more affected by group size than by the number of groups.

(6) The precisions of variance components estimates, in contrast to that of the macro parameters, are directly related to intraclass correlation coefficient. (7) Departures of empirical type I error rates from nominal alpha for tests of macro parameters are typically within 99% confidence intervals. When outside the probability intervals, empirical significance levels are all liberal. No pattern developed between empirical type I error rates and number of groups, group size or effect size. (8) For all macro parameters, power increases as total sample size, number of groups, group size, or effect size increase. However, group size has a consistent determinative effect on power over the number of groups.

To Mohammad and Yashaar

Acknowledgements

I would like to take this opportunity to thank my committee chairperson, counselor and friend Dr. Stephen W. Raudenbush for his invaluable support, insightful comments and understanding. Working with him contributed greatly to my professional development.

I would also like to thank Dr. Richard F. Houang who has been a constant source of inspiration, encouragement and genuine support throughout my graduate studies, as well as for his intellectual persuasion throughout this research.

I wish to express my appreciation to the rest of my committee, Drs. Dennis Gilliland, William H. Schmidt, and Robert E. Floden for reviewing my work and providing suggestions for improvement. I would further like to take this opportuniy to acknowledge the support I received from the Spencer Foundation.

Most importantly, my deepest gratitude goes to my dearest friend and husband, without whose understanding, patience and moral support this work certainly would not have been completed, the person who always knew I could do it, Dr. Mohammad Ali Chaichian.

I wish to express my appreciation to my parents who have been a source of love and strength from before my time at Michigan State University. Last but not the least, my

deepest appreciation goes to my son Yashaar for his patience and understanding while I worked on my dissertaion.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

STATEMENT OF THE PROBLEM

Science's main job is to "explain" natural phenomena by discovering and studying the relations among variables. In the behavioral sciences, variability is itself a phenomenon of great scientific curiosity and interest. In their attempts to explain the variability of a phenomenon of interest (often referred to as the dependent variable), scientists study its relations or covariations with other variables (referred to as the independent variables). Educational researchers seek to explain the variance of school achievement by studying its relations with intelligence, aptitude, social class, race, home background, school atmosphere, teacher characteristics, and so on. Various analytic techniques have been developed for the purpose of studying relations between independent variables and dependent variables, or the effects of the former on the latter (Pedhazur, 1973).

Perhaps the most powerful method of doing this is the regression analysis, whose simplest form is one in which the effect of an independent variable on a dependent variable is being studied (Pedhazur, 1973). Under this simple conception the two parameters of interest are the slope and intercept

which are usually called regression coefficients. The test statistic used for either of the regression coefficients is the Z-test (or t-test if the sampling variance, σ_e^2 , is replaced by its unbiased estimator). So long as we deal with a situation where the variables have the same level of aggregation (e.g., both at individual or group level) and where our measurement processes are assumed to be error free, there is no real drawback to this approach.

But in educational research many, if not most, data have multilevel characteristics. For example, students are nested within classrooms, and classrooms are nested within grade levels which are themselves nested within schools, districts, or program sites. Thus, we can have variables of different levels, describing students, classes, schools and so on. Variables such as family background, prior achievement, parental educational level, and the like are individual (or micro) variables identified with students and variables such as whether the school is public or private are group (or macro) variables. In a multilevel problem we want to investigate the relations between variables at different levels of hierarchy as well as interactions across levels.

There has been a great surge of interest in educational statistics over the past decade to search for appropriate statistical methods for hierarchical, multilevel data. As a result of this search, a general approach to the problem of multilevel data which is referred to as hierarchical linear

models (HLM) by Sternio (1981) has emerged.

The basic idea of a hierarchical linear model is fairly simple. When data are available at two levels of aggregation, for example, on students and the schools to which they belong, the model is specified in two sets of equations: one within schools, and one between schools. The within-school model is defined separately for each school with student level predictors and a student level outcome variable. This is a familiar linear regression model with one major exception; the within-school parameters, regression coefficients, are allowed to vary randomly across schools. This conception poses a second or between-school model. The between-school model then regresses the within-school regression coefficients on to the school level predictors.

Two sets of parameters evolve from this formulation: micro parameters, or random effects, and macro parameters, or fixed effects. Research interest has focused on estimation of both micro and macro parameters. As Raudenbush (1988: 87) has pointed out:

> "Two fundamentally different types of problems have motivated the development of these HLM models. In the first type, interest focuses on the micro parameters or random effects. One seeks to estimate, for instance, a regression equation for a particular school, the effect size for a particular study, or the growth rate of a particular child when the data available for that school, study, or child are sparse. The empirical Bayes approach strengthens estimation for each unit by utilizing data from many similar units: schools, studies, or children. In the second type of application, attention focuses on the macro parameters, or fixed effects. One asks why some kinds of schools have smaller regression slopes than others, why some studies report larger

effects than others, and why some children grow faster than others."

The conception that "micro" parameters vary randomly across the population of groups as a function of "macro" parameters not only justifies the "slopes as outcomes" idea (Burstein, 1980), but also introduces a new source of variation in micro parameters referred to as random effect variance or parameter variance. This is the variance among the micro parameters themselves which is distinguished from the sampling variance resulting from using a sample within each macro unit to estimate these parameters.

The new advances in analyzing multilevel data have evolved from statistical theory stimulated by the seminal contributions of Lindley and Smith (1972), Novick, Jackson & Thayer (1972), and Smith (1973), who developed Bayesian estimation procedures for hierarchicaly structured data . When the variance components (i.e., sampling and parameter variance) are known, estimates for the micro and macro parameters can be derived from alternative estimation theories: least squares, Bayesian, and maximum likelihood (see for example Raudenbush, 1984). The crucial difference between Bayesian/ maximum likelihood approach and least squares approach is the difference in assumptions.

In most applications, however, these variance components have to be estimated. Unfortunately, no simple closed-form estimate is available. However, a variety of numerical approaches to maximum likelihood estimation of covariance

components are available among which EM algorithm (Dempster, Laird & Rubin, 1977) is especially conceptually appealing.

The EM algorithm produces maximum likelihood estimates for variance components with known large sample properties. That is, the estimates are consistent, asymptotically efficient with known large sample normal distributions. The fact that the sampling distributions are known becomes especially important when inferences are to be made based on the parameter estimates. The test statistic for a macro regression coefficient is a Z-test (asymptotic Z-test or t-test if the variance components are replaced by their maximum likelihood estimates). But before asymptotic results become exact, the number of levels of each random factor must increase to infinity (Miller, 1977). That is, for example, both the number of schools (call it K) and the number of pupils (call it n) within each school must approach infinity.

At present little is known about the small sample behavior of these parameter estimates. To date it is not clear how large n and K have to be in order for estimates and their standard errors to become acceptable, thus justifying the use of large sample theory.

The goal of this research is to understand the properties of maximum likelihood estimates obtained from small and moderate samples, and to evaluate their implications for research design. Because anlytic study of these properties becomes intractable in the case of unknown variances and

covariances, empirical studies are needed. Clearly to gain a comprehensive understanding of the inferential strength of hierarchical linear model, and to understand the small sample properties of its parameter estimates, many simulation studies are needed. In other words, alternative HLM methods with different model specifications and assumptions, or at least the most interesting and realistic ones, have to be studied.

This research will take the initial step and will address these issues by considering the two-stage standardized hierarchical linear model. Specifically, this research through simulated data generated for different values of K and n, will investigate the effects of variance estimation via EM algorithm on inferences about parameters at the second stage of the hierarchy, that is, about the macro or fixed effects.

The following chapters will review the literature with respect to statistical approaches to multilevel data, discuss maximum likelihood parameter estimation, present the method used for investigating the small sample properties of parameter estimates, and provide results and discussion of their implications for research design.

CHAPTER II

REVIEW OF THE LITERATURE

A long-standing problem associated with educational research has been the failure of many quantitative studies to attend to the complexity of data usually produced by hierarchical, multilevel educational field research (Cronbach, 1976; Haney, 1980; Burstein, 1980; Cooley, Bond & Mao, 1981; Rogosa, 1978).

Cronbach (1976) remarked that the majority of studies of educational effects carried out until 1976 conceal more than they reveal, and that "the established methods have generated false conclusion in many studies." (P. 1)

Uni-level Techniques

Traditionally, statistical approaches have attempted to adopt uni-level techniques to multilevel situations. This can often be done by using aggregation or disaggregation. A student (micro) variable, such as intelligence, can be aggregated to school level by assigning to a school the average intelligence of its students. A school (macro) variable, such as whether it is public or private, can be disaggregated to the student level by assigning to each student the type of school. But as de Leeuw and Kreft (1986) pointed out "the operations of aggregation and disaggregation are highly nontrivial, both from the methodological and

from the statistical point of view". Conceptually, by aggregating, a change in the meaning of the variables occurs. Statistically this means we are ignoring all within-school variation which sometimes results in dramatic increase in the correlation between aggregated variables. Robinson (1950) showed that not only does the correlation fluctuate as a function of grouping but that the sign may even be different at different levels. As a result we no longer can make inferences on the student level without committing the `ecological fallacy' (Alker, 1969; Cronbach and Webb, 1975; Hannan, 1971; Robinson, 1950). This refers to the practice of interpreting correlations between aggregated variables as if they were correlations between variables measured on individuals (i.e., cross level inference). This is the most commonly cited flaw in any early methodological treatments of hierarchical data.

On the other hand if we disaggregate, we have to take into account the fact that students within the same school do not respond independently to school level variable. But the traditional linear models require the assumption that subjects respond independently to educational programs. Also by ignoring the nested structure of data we will misestimate the precision of parameter estimates, resulting in serious inferential problem (Aitkin, Anderson & Hinde, 1981; Knapp, 1977; Walsh, 1947).

In the late 1960s and early 1970s, the topic of aggregation and proper choice of analytic units (using the

student versus using the group) gained popularity in educational field research (see Burstein, 1980 for a review of these issues). This increased interest may be viewed as a natural by-product of the then growing emphasis in educational research on evaluation of social and educational programs; evaluations that had to be designed and analyzed in such a way as to take into account the ever-present natural hierarchy found in all school systems. With the awareness that students within a class and teachers within a school cannot be considered truly independent and that responses to treatment may rightfully vary dramatically across groups, came an increased interest in how to deal with non-independence and differential effects for distinct population groups often labeled aptitude-treatment interactions (Cronbach and Webb, 1975).

Up to the early 1970s, then, considerable research had been done on the effects of aggregation on bias and efficiency under various grouping strategies. However, little of this research was grounded in practical applications.

As Burnstein (1980) in his concluding remarks on choice of units of analysis points out, if the goal is to learn something about the effect of educational process on student achievements, the "discussion, about the choice of an appropriate unit are simply unnecessary digressions" (p. 196). The emphasis should be on choosing an appropriate analytical model that accounts for the relationship among

variables observed at both levels of aggregation. Rogosa (1978: 83) remarked, "no one level is uniquely responsible for the delivery of and the response to educational programs.... confining substantive questions to any one level of analysis is unlikely to be a productive research strategy". Burstein and his associates (Burstein, Linn & Capell, 1978; Burstein, Miller & Linn, 1979; Burstein, 1980) argue that when relationships between the dependent and independent variables are different in different groups, single level analyses at either the individual or aggregate level will produce misleading results. For example, conducting analyses at the individual level without regard for group membership might lead to spurious null effects or to spurious large effects; in either case, the actual effects will only be uncovered through within-group analysis. Thus, these researchers advocate conducting selected within-class analyses and using the results of these regressions in aggregate level analyses. Cooley et al. (1981) reached the same conclusions and pointed out : "We must not ignore the possibility of variation among groups (e.g., classrooms or schools) in estimating a variable's effect. Examining this variation can reveal grouping effects or specification error, ignoring it will conceal them." (p. 74)

Such criticisms of single level analyses suggest that multilevel approaches are needed in many settings. Such models would aim simultaneously to discover: 1) what is happening within macro units; 2) what differences there are

between macro units; and 3) how those differences influence the quality of what is going on within the macro units. To be valid, statistical analyses must account simultaneously for effects at both levels.

Multilevel Techniques With Random Intercepts and Fixed Slopes

In the mid-seventies, the problem of aggregation bias was resolved by analyzing multilevel data with multilevel techniques. Some of these alternative analytical strategies are: the separate between-group and pooled within-group analyses as suggested by Cronbach (1976), and Cronbach and Webb (1975); a two-stage hierarchical analysis proposed by Keesling and Wiley (1974), and Wiley (1976); and a "full model" analysis suggested by Keesling (1977). All three strategies obtain their estimators through ordinary least squares (OLS) technique, but they differ in the approach by which the estimators are obtained. Notice that in the case of random intercepts, OLS is an appropriate estimation method only when balanced designs are considered.

Schmidt and Houang (1983) compare and contrast these three approaches with respect to parameter estimation. They concluded that these strategies differ in the way that the relationship between the between-group effects and the within-group effects are conceptualized. Analytically they showed that all three procedures give the same estimate for the within-group regression coefficient. With respect to the between-group regression coefficient, the estimate

obtained by Cronbach's approach is different but related to that of Keesling's. As far as the third approach is concerned (i.e., Keesling and Wiley's two-stage analysis), no estimate for the between group regression coefficient is available in this case. These differences reflect different conceptualizations in the three strategies. That is, where in the Cronbach's and Keesling's approaches the individual level variables are conceptualized to have direct impact on the outcome variables, in the Keesling and Wiley's their influence are indirect and mediated through other group level variables. This difference in conceptualization of the situation sets a criterion for choosing among these three multilevel techniques (Schmidt and Houang, 1983).

The model proposed by Wisenbaker and Schmidt (1979) may be viewed as the extension of these multilevel techniques to their multi-variate form. The application of a "components of covariance structure" method (Schmidt, 1969) to the proposed, multivariate random effects model (with random intercepts and fixed structural parameters) allow the simultaneous estimation of between-group and within-group effects and their standard errors via maximum likelihood procedure. The model potentially permits different specifications for the relations at two levels of hierarchy.

Although these analytical strategies accounted for aggregation bias, however, the other technical problem, i.e., misestimated precision, remained unresolved. The problem is that these methods allow for random intercepts,

but they assume constant within-group slopes. The technical consequences of ignoring slope heterogeneity when in fact it exists are inefficient estimation of regression coefficients and negatively biased standard errors of regression coefficient, which inflate type I error rate.

Cronbach (1976) cites three sources of variation in within-class slopes: 1) sampling variability and stability problems due to small class sizes when the process operating in the classes are basically the same; 2) differences in the selection factors operating to form the classes; and 3) differences in causal processes going on in the classrooms. If we can rule out chance effects and different selection rules as reasonable explanations, the variation in within-class slopes become a potent source of information to researchers and policy makers.

Tate and Wongbundhit (1983) argued that random coefficient models with random slopes and intercepts are more appropriate than random coefficient models with only a random intercept for multilevel analysis in educational research. De Leeuw and Kreft (1986) further added that random coefficient models are more general and that fixed constants are special random variables. They argued that "whether something is random or fixed should be decided by considering what would happen if we replicate the experiment. Would it be realistic to suppose that regression coefficients stayed the same under replications? If not, then random coefficients are appropriate" (p. 59).

Boyd and Iversen (1979) discussed a "separate equations" approach in which both intercept and slope are allowed to be random. But their estimation procedure is ordinary unweighted least squares for both sets of coefficients, which ignores the information provided by the random coefficient model.

One class of multilevel approaches in which random coefficients are estimated (Burstein and Miller, 1980; Cooley et al., 1981), first estimates relationships within each school; then these regression coefficients serve as outcome variables for an assessment of the importance of school policies and practices. Again, this approach is not free from problems. Raudenbush and Bryk (1986) discussed some of the technical difficulties associated with slopes-as-outcomes approach. Among these problems are weak statistical power to detect real differences in slopes, the need for a multivariate formulation so that several regression coefficients per unit can be studied, and the need for a statistical model which matches the complexity of hierarchical, multilevel character of most educational field research data.

Additional problems with Cooley et al.'s proposed model is that all random variation in the effects of macro units is assumed to be explained by the predictors included in the model, so that the only unexplained variation results from sampling of micro units (i.e., unsystematic or sampling variation). In traditional analysis of variance terms, such

a model is a fixed effects model. Assuming the model is completely specified, there is no drawback to this approach. However, when the model specification is incomplete, which will commonly be the case, the parameter estimates of regression coefficients and their standard errors are untrustworthy.

The Multilevel Technique With Random Intercepts and Random Slopes:HLM

A general approach to the problem of multilevel data (Aitken and Longford, 1986; de Leeuw and Kreft, 1986; Goldstein, 1986; Mason, Wong & Entwisle, 1984; and Raudenbush and Bryk, 1986) incorporates the idea of "slopes as outcomes" without its various deficiencies. This general approach with random effects at each sampling level has been proposed under a variety of names: variance component models (Harville, 1977), mixed model ANOVA (Elston and Grizzle, 1962), regression with random coefficients (Rao, 1972; Swamy, 1973; Rosenberg, 1973; and Dielman, 1983), Bayesian estimation for linear models (Lindley and Smith, 1972; Smith, 1973,Dempster, Rubin & Tsutakawa, 1981; and Morris, 1983), multilevel linear models (Mason et al., 1984), mixed linear models (Goldstein, 1986), and hierarchical linear models (HLM) (Sternio, Weisberg & Bryk, 1983) have all been used. The present study employs the term hierarchical linear models, labeled HLM for convenience.

The HLM has a hierarchical structure in the sense that parameters at a lower level of aggregation (i.e., micro

parameters) are assumed to vary over a population of groups as a function of the parameters at the next higher level (i.e., macro level). Micro parameters may be as diverse as means, proportions, variances, linear regression coefficients and logit linear regression coefficients (see Raudenbush, 1988). Through such models, it is possible to assess the strength of relationship between macro predictors and micro parameters. This quality along with the "slopes as outcomes" idea enables investigators to go beyond traditional questions (e.g., why do more schools have higher achievement than others ?) and ask more fundamental questions about why structural relationships vary across groups. This class of questions (e.g., why is the effect of social class or race stronger in some schools than others ?) reflect the "slopes as outcomes" conceptualization popularized by Burstein (1980). The HLM model identifies both slope and intercept heterogeneity and tries to explain them via related macro predictors.

Not only do such models enrich the class of research questions asked about educational effects occuring within and between educational units, they solve problems of aggregation bias and misestimated precision long associated with multilevel data.

Estimators of micro and macro parameters are available through empirical Bayes methods. The empirical Bayes estimates of the micro parameters (also called shrinkage or Stein estimators) provide an improvement over the least

squares estimators. This improvement is most pronounced when some or all groups have sparse data and when there is heterogeneity among micro parameters, some of which can be explained by group characteristics.

Estimation of the micro parameters can be improved by shrinkage of least squares estimates around a grand mean (known as "unconditional shrinkage") in the first situation and by shrinkage toward a conditional expectation (known as "conditional shrinkage") in the second situation.

The empirical Bayes approach also yields estimates for macro parameters. This estimator, which is recognizable as the generalized least squares estimator, weights each OLS estimate of micro parameters proportional to its precision.

Estimation of macro parameters are of great importance not primarily because these improve estimation of micro parameters, but because it enriches the class of research questions asked about educational effects which goes far beyond what was plausible prior to the advent of HLM models.

Research interest has focused on estimation of both micro parameters and macro parameters each addressing fundamentally different type of questions. Studies with the goal of improving micro estimators (by either conditional or unconditional shrinkage) include Laird and Ware, 1982; Raudenbush and Bryk, 1985; and Sternio, et al., 1983 for the first type of shrinkage and Braun, Jones & Rubin, 1983; Der Simonian and Laird, 1983; Novick, et al., 1972; Novick and Jackson, 1974; Rubin, 1980 and 1981; and Shigemasu, 1976 for

the second type of shrinkage. However, numerous investigators have recently found that the macro parameters themselves may be of greater interest (Aitkin and Longford, 1986; Aitkin, et al., 1981; de Leeuw and Kreft, 1986; Goldstein, 1986; Laird and Ware, 1982; Lee, 1986; Mason, et al., 1984; Raudenbush and Bryk, 1985 and 1986; and Sternio, et al., 1983).

The HLM model has broad applicability in educational research. The study of individual growth (Laird and Ware, 1982; Sternio, et al., 1983; Bock, 1983), the measurement of change (Bryk and Raudenbush, 1987); contextual effects in cross-national fertility research (Mason et al., 1984), and research synthesis or "meta analysis" (Raudenbush and Bryk, 1985) are examples of HLM's broad applicability.

The major problem with this development is the mathematical complexity of Bayesian covariance components estimation. Fortunately, a variety of numerical approaches to maximum likelihood estimation of covariance components are now available.

Estimation of Dispersion Matrices:

Estimation of dispersion matrices in multilevel linear models with fixed and random effects (i.e., mixed models) can be complex, particularly in an unbalanced case. The traditional 'ANOVA' approach is essentially the only method in use for balanced data. This method consists of equating the observed sums of squares and cross-products matrices to their expected values. For unbalanced data, the 'ANOVA'

approach leads to biased estimators of variance components. Henderson (1953) developed analogous techniques to correct this deficiency. Searle (1968, 1971a, and 1971b) gives excellent descriptions of Henderson's methods and indicates various generalizations. One problem with Henderson's method for estimating variance and covariance components is that the methods are not necessarily well defined. Moreover, except for balanced data cases, little is known about the properties of the Henderson estimators, other than that they are unbiased and translation invariant. It is known that, at least in particular cases, there are biased estimators that have uniformly (assuming normality) smaller MSE's than the Henderson estimators (see Klotz, Milton and Zacks, 1969). The discovery by Seely (1975) and by Olsen, Seely, and Birkes (1976) proved that, at least in the case of most unbalanced mixed or random effects models having one random factor, there exist estimators that have uniformly smaller variance than the Henderson estimators. These locally best estimators are related closely to maximum likelihood estimators (Hocking and Kutner, 1975).

Maximum likelihood and related procedures, which are reviewed by Harville (1977), have received increased attention in the past ten years. However, maximum likelihood approach has been somewhat ignored by practitioners because of computational complexities and because it takes no account of the loss in degrees of freedom (df) from the estimation of fixed effects, leading in some instances to

large biases and large mean squared errors (Patterson and Thompson, 1974). Improved computational procedures are now available, and Patterson and Thompson (1971, 1974) have devised a modified ML approach known as 'resticted maximum likelihood', that adjusts automatically for losses in df. As Harville (1977:320) states:

> " Certain deficiencies of various other methods are not shared by maximum likelihood. In particular, the maximum likelihood approach is 'always' well defined, even for the many useful generalizations of the ordinary ANOVA models, and, with maximum likelihood, nonnegativity constraints on the variance components or other constraints on the parameter space cause no conceptual difficulties. Moreover, the maximum likelihood estimates and the information matrix for a given parameterization of the model can be obtained readily from those for any other parameterization".

Asymptotic Properties of Maximum Likelihood Estimates

The attractive features of maximum likelihood estimates of variance-covariance components, discussed by Harville (1977), are important. The maximum likelihood are functions of sufficient statistic and are consistent, i.e., they converge to the population values as the sample size becomes indefinitely large. Their joint distribution is approximated by the multivariate normal distribution with mean equal to the population value and variance-covariance matrix equal to the negative inverse of the matrix of second derivative of the likelihood function. Moreover, the maximum likelihood estimators are said to be asymptotically efficient (in the sense described by Miller, 1973 and 1977) attaining the Cramer-Rao lower bound for the covariance matrix under mild regularity condition . There is, however, no guarantee or

unbiasedness or efficiency in small samples.

In order to obtain asymptotic results in the mixed model, the number of levels of each random factor must increase to infinity. More often in the analysis of variance a conceptual sequence of experiments with the number of levels of each of the random factors increasing to infinity is considered. Hartley and Rao (1967) were the first to attempt an asymptotic theory that would be truly appropriate for the more complicated of the ordinary ANOVA models. They proved that under certain restrictions the estimates were consistent and asymptotically normal as the size of the experimental design increased. However, one of their assumptions is that the number of observations at any level of any factor must remain less than some fixed constant for all designs in the sequence. This assumption eliminates many crossed designs where the number of observations at a given level of one factor is proportional to the number of levels of another factor.

An alternative way of obtaining asymptotic results in the mixed model is by considering repetitions of a given experiment. Anderson (1969, 1971) considered maximum likelihood estimates in a more general class of models (multivariate models where the covariance matrix has linear structure) and proposed a different solution; he proved that the estimates were consistent and asymptotically normall as the entire design was repeated. Miller (1973) developed an asymptotic theory for the ordinary ANOVA models which, while

it is similar to that presented by Hartley and Rao (1967), it does not exclude any cases of real interest. He considered asymptotic properties of the maximum likelihood estimates for a large class of design sequences whose size increases to infinity; this class of design sequences contains all sequences treated by Hartley and Rao and most sequences which could occur in practice. In other words he took the basic model of Hartley and Rao, rewrote it in the form used by Anderson and proved consistency and asymptotic normality of the estimates in the model.

Raudenbush (1988) in his paper entitled "Educational Application of Hierarchical Linear Models: A Review" provides a comprehensive review of HLM model with respect to estimation theory and application. In his concluding remarks he states, "despite the clear potential of such models, important questions about their statistical properties remained unanswered. The questions concern small sample properties, implications for research design and robustness of violations of assumptions" (p. 111). This research will take the initial step and will address questions about small sample properties of the estimators and their implications for research design by considering a two-stage standardized hierarchical linear model.

CHAPTER III

TWO STAGE HIERARCHICAL LINEAR MODEL (HLM)

In this chapter, a mathematical model for the general two-stage hierarchical linear model (HLM) is presented. This is followed by the description of parameter estimation when variance components are known and when they are unknown. Then the logic of EM algorithm along with the steps involved for the implementation is discussed. Finally, the effects of estimating variance components on macro or fixed parameters is described.

For reasons of simplicity and clarity, a two-stage hierarchical linear model is considered although the statistical theory permits more (see Goldstein, 1986). The basic idea of HLM is reasonably simple. We begin by supposing that the researcher has data at two levels of aggregation, for example, on students and the schools to which they belong. The model is specified in two sets of equations: one within schools, and one between schools. Our fundamental assumption is that the outcome variable in some way depends on the student level predictors and that the micro regression coefficients may vary systematically as a function of the school level predictors. The within school model is defined separately for each school. This is a

familiar linear regression model, with student level predictors, and with a student level outcome variable. The between-school model then regresses the within-school regression coefficients on to the school level predictors. The present study restricts attention to the case in which variation in the outcome variable, Y, is to be explained as a function of one student (micro) level predictor, X, and one school (macro) level predictor, W (theoretically there is no limit as to the number of Y, X and W). In this case of a simple univariate regression model the within-school model (or micro model) becomes

$$Y_{ij} = \mu_j + \beta_j X_{ij} + R_{ij} \qquad (3.1)$$

and

$$R_{ij} \sim N(0 , \sigma_j^2)$$

where

Y_{ij} is the outcome score for student i in school j; where j = 1,..., n

μ_j and β_j are the micro level regression coefficients within school j;

X_{ij} is the micro level predictor for student i in school j; and

R_{ij} represents random error, assumed independently normally distributed with zero mean and variance σ_j^2.

By centering the micro level predictor around its respective group mean, $X_{\cdot j}$, μ_j represents the mean on the

outcome variable in school j. Equation (3.1) is a standard linear "full rank" regression model with one major exception; the within-school parameters, μ_j and β_j are allowed to vary randomly across schools. This conception poses a second or between-school model.

The between-school model (or macro model) may be either unconditional (involving no macro level predictors) or conditional (involving macro level predictors). The unconditional model is:

$$\mu_j = \bar{\mu} + U_{0j}, \tag{3.2}$$

$$\beta_j = \bar{\beta} + U_{1j}, \tag{3.3}$$

and

$$U_{0j} \sim N(0, \tau_\mu),$$

$$U_{1j} \sim N(0, \tau_\beta),$$

$$\text{cov}(U_{0j}, U_{1j}) = \tau_{\mu\beta},$$

that is, μ_j and β_j are viewed as a functions of their respective grand mean across all schools plus random error. Under this simple model, τ_μ and τ_β represent the parameter variances in U_{0j} and U_{1j} respectively. That is, they signify the variability in the true intercept and slope across the population of schools, and that $\tau_{\mu\beta}$ signifies the covariance between them. Treating W as potential determinant of μ_j and β_j, leads to the following conditional between-school model:

$$\mu_j = \gamma_{00} + \gamma_{01} W_j + U_{0j} \tag{3.4}$$

$$\beta_j = \gamma_{10} + \gamma_{11} W_j + U_{1j} \qquad (3.5)$$

and

$$U_{0j} \sim N (0 , \tau_{\mu|W})$$

$$U_{1j} \sim N (0 , \tau_{\beta|W})$$

$$\text{cov} (U_{0j} , U_{1j}) = \tau_{\mu\beta|W}$$

where $\tau_{\mu|W}$ and $\tau_{\beta|W}$ are the conditional parameter variances in U_{0j} and U_{1j} respectively, and $\tau_{\mu\beta|W}$ is their conditional covariance. The micro errors are assumed independent of the macro errors. Equations (3.4) and (3.5) represent the effects of macro predictor W on the two micro parameters, μ_j and β_j. These two equations combined with equation (3.1) define a multilevel model that can be written equivalently as a single equation by substituting (3.4) and (3.5) into (3.1):

$$Y_{ij} = \gamma_{00} + \gamma_{01} W_j + \gamma_{10} X_{ij} + \gamma_{11} X_{ij} W_j + (U_{0j} + X_{ij} U_{1j} + R_{ij}) \qquad (3.6)$$

The brackets in equation (3.6) enclose error terms that complicate the expression considerably, as they do its estimation. The presence of macro error terms in (3.4) and (3.5) make (3.6) a mixed model, because it contains fixed coefficients (the γ's) and random coefficients (the U's).

The model shown in equation (3.6) is quite general in that a number of familiar models can be derived from it. If the macro errors are suppressed, the hierarchical linear model (3.6) becomes equivalent to an ordinary regression

model (or fixed effect specification) that includes student level variable X_{ij} , school level variable, W_j , and their interaction effect, $X_{ij}W_j$, and its estimation poses no special problem. Under this model we assume that all of the variation in the micro parameters, μ_j and β_j , has been perfectly explained by knowledge of the macro level variable, W_j , whereas equation (3.6) allows for error. When random effects remain (i.e., U_{0j} and/or U_{1j} are not equal to zero), application of ordinary least squares to (3.6) is inefficient, and the estimated standard errors are too small. Another model that has received some attention is "random intercept regression model". This model considers the within-school intercepts, μ_j , as random, but the regression slopes, β_j , as fixed. Some variant of this model has been employed by Aitkin et al. (1981), Cronbach (1976), Keesling (1977), Keesling and Wiley (1974), and Wisenbacker and Schmidt (1979). There are hypothesis tests in each case to decide whether or not it is justifiable to make these simplifications (Raudenbush and Bryk, 1986). Mason et al. (1984) provide a detailed discussion of the relationship between the general hierarchiacl linear model (3.6) and other simplified sub-models of potential interest that can be derived from it.

Estimation Under Known Variance Components

Estimates for the parameters in HLM models assuming known variance components can be derived from alternative estimation theories: least squares, Bayesian and maximum

likelihood (see for example Raudenbush, 1984). Using matrix notations to generalize the model, equation (3.1) becomes;

$$Y_j = X_j \theta_j + R_j, \qquad (3.7)$$

and

$$R_j \sim N(0, \Sigma_j),$$

where

$$\Sigma_j = \sigma_j^2 I_{n_j},$$

$$Y_j = \begin{vmatrix} Y_1 \\ \vdots \\ Y_{n_j} \end{vmatrix}, \quad \theta_j = \begin{vmatrix} \mu_j \\ \beta_j \end{vmatrix}, \text{ and } R_j = \begin{vmatrix} R_1 \\ \vdots \\ R_{n_j} \end{vmatrix},$$

and equations (3.4) and (3.5) will be reduced into a single equation of the form,

$$\theta_j = \mu_\theta + \mu_j, \qquad (3.8)$$

and

$$U_j \sim N(0, \tau),$$

where

$$\tau = \begin{vmatrix} \tau_\mu & \tau_{\mu\beta} \\ \tau_{\mu\beta} & \tau_\beta \end{vmatrix}$$

$$\mu_\theta = \begin{vmatrix} \bar{\mu} \\ \bar{\beta} \end{vmatrix}, \quad \text{and} \quad U_j = \begin{vmatrix} U_{0j} \\ U_{ij} \end{vmatrix}.$$

Under the Bayesian approach, assuming variance components are known, the minimum mean squared error point estimators for micro and macro parameters are;

$$\theta_j^* = \lambda_j \hat{\theta}_j + (I - \lambda_j)\, \mu_\theta^*, \qquad (3.9)$$

and

$$\mu_\theta^* = (\Sigma\lambda_j)^{-1} \quad (\Sigma\lambda_j \hat{\theta}_j) , \qquad (3.10)$$

where

$$\lambda_j = \tau \ (\tau + \ V_j \ )^{-1} ,$$

$\hat{\theta}_j$ is the ordinary least squares estimate of θ_j for each school with sampling error of

$$\text{var} \ (\ \hat{\theta}_j \ \Big| \ \theta_j \ ) = \sigma_j^2 \ (\ X_j' \ X_j \ )^{-1} = V_j \ ,$$

and λ_j represents a "multivariate ratio" of the true parameter variance in θ_j to the total observed variance in $\hat{\theta}_j$. This ratio signifies the reliability of $\hat{\theta}_j$ as an estimator of school j's slope. It follows that $\hat{\theta}_j = (X_j'X_j)^{-1}X_j'Y_j$ is normally distributed with a mean of μ_θ and variance $V_j + \tau$ i.e.,

$$\text{var} \ (\ \hat{\theta}_j \ ) = \text{var} \ (\ \hat{\theta}_j \Big| \ \theta_j \ ) + \text{var} \ (\ \theta_j \ ) = V_j + \tau = \Delta_j \ .$$

The empirical Bayes estimator, μ_θ^* is a generalized least squares estimate of θ , where the outcome vector $\hat{\theta}_j$ (i.e., OLS estimates of micro regression coefficients) is weighted by its precision. The empirical Bayes estimator θ_j^* is a weighted combination: first of $\hat{\theta}_j$, the OLS slopes derived for each school based only on the student data from that school; and second, from the estimated mean slope μ_θ^* , for the population of schools. That is, θ_j^* is a vector the elements of which are somewhere between the elements contained in $\hat{\theta}_j$ derived entirely from within the macro unit j, and the elements μ_θ^* the estimated mean vector for the

entire sample.

The properties of these estimators are reviewed by Efron and Morris (1975) and Morris (1983). Such estimators are conditionally biased, i.e., the bias is largest for θ_j values far from the average. However, in general θ_j^* is a more precise estimator (i.e., it has smaller expected mean squared error) than $\hat{\theta}_j$ its OLS counterpart.

Sternio (1981) reasoned, however, that the precision of θ_j^* could be improved even further by shrinking estimates $\hat{\theta}_j$, not toward a grand mean μ_θ^*, but toward a conditional mean $W_j\gamma^*$. This is obtained by regressing θ_j onto a macro parameter W as follows,

$$\theta_j = W_j\gamma + U_j \ , \qquad (3.11)$$

where

$$\gamma = \begin{vmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{vmatrix} \ , \quad W_j = \begin{vmatrix} 1 & W_j & 0 & 0 \\ 0 & 0 & 1 & W_j \end{vmatrix} \ .$$

Under this formulation the empirical Bayes estimators or equivalently the posterior means for micro and macro parameters are:

$$\theta_j^* = \lambda_j \hat{\theta}_j + (I - \lambda_j) W_j\gamma^* \ , \qquad (3.12)$$

$$\gamma^* = (\Sigma W_j' \Delta_j^{-1} W_j)^{-1} \Sigma W_j' \Delta_j^{-1} \hat{\theta}_j \ , \qquad (3.13)$$

where

$$\lambda_j = (\tau|W) (\tau|W + V_j)^{-1} \ .$$

These results generalize to the case of multiple X's and multiple W's. The posterior dispersion of θ and γ are given in equations (3.14) and (3.15) respectively (Raudenbush, 1988: 91).

$$D^*_{\theta_j} = \lambda_j V_j + (I - \lambda_j) S_j (I - \lambda_j)' \quad , \qquad (3.14)$$

where

$$S_j = W_j (\Sigma W_j' \Delta_j^{-1} W_j)^{-1} W_j' ,$$

and

$$D^*_{\gamma} = (\Sigma W_j' \Delta_j^{-1} W_j)^{-1} \quad . \qquad (3.15)$$

It is worth noting that the crucial difference between the three alternative estimation theories: least squares, Bayesian and maximum likelihood is the difference in assumptions. With regard to θ , the Bayesian and maximum likelihood method lead to identical result but different from least squares. This is because the least squares method makes no assumptions about the prior distribution of θ_j . On the other hand, both Bayes and maximum likelihood assume normality of the θ_j in order to derive θ^*_j . With regard to γ all three approaches effectively assume no prior distribution and therefore produce identical results (Raudenbush, 1984).

Estimation Under Unknown Variance Components

So far we have assumed that the variance components are known. In most applications however, these will not be given

and have to be estimated. For balanced data it has been common practice to equate the observed sums of squares and cross-products matrices to their expected values (called "ANOVA" approach). Estimating variance components from unbalanced data is not as straight forward as from balanced data. Henderson (1953) developed analogous techniques dealing with variance component estimation from unbalanced data. However, his method is computationally cumbersome when a mixed model is assumed and when the number of classes is large. Searle (1971a, Chapter 10) discussed problems with the ANOVA approach when applied to unbalanced data.

As can be seen in the survey article by Searle (1971b) there are many approaches to variance components estimation from unbalanced data; many of them of a rather specialized nature and many which depend on some form of balance or symmetry in the problems addressed. The need for general procedures for handling unbalanced problems is quite well known to the statisticians.

A complete Bayesian analysis can be performed by specifying a joint prior distribution for all the parameters involved (θ , γ , τ and V in our case), combined with the likelihood function for regression coefficients (here θ and γ) in order to obtain a joint posterior distribution for the four parameters. This distribution then has to be integrated with respect to the variance-covariance components (here τ and V), thus removing the nuisance parameters, so that the posterior distribution of θ and γ can be

calculated (Lindley and Smith, 1972). While theoretically satisfying, this approach is computationally complex.

As a practical alternative Sternio, et al. (1983) followed the general strategy of Dempster, et al. (1981) and developed an empirical Bayes analysis. The empirical Bayes approach consists of first deriving Bayesian estimates based on known variances and then substituting maximum likelihood estimates for the unknown variances and covariances in the estimation formulas. Similarly, they generated maximum likelihood estimates for unknown variance-covariance components via EM akgorithm (Dempster, et al., 1977), and then replaced the true parameter values in their model by these estimates.

As Harville (1977: 320) points out, ".... except in relatively simple settings (cases), the computation of maximum likelihood estimates requires the numerical solution of a constrained non-linear optimization problem". For unbalanced data maximum likelihood estimates of variance components are not available in closed form and one has to resort to iterative solutions to obtain them.

A variety of numerical approaches to maximum likelihood estimation of variance-covariance components are available. Among them EM algorithm is specially gaining prominence. Dempster, et al. (1977), review many areas where the EM algorithm has successfully been applied, or has potential applications. These include missing value situations, aplli-cations to grouped, censored or truncated data, variance

component estimations, iteratively reweighted least squares, fixed mixture models, hyperparameter estimation and factor analysis. They also derive theorems showing the monotonic behavior of the likelihood function and the convergence of the algorithm. Some of the applications include Aitken, et al. (1981), Dempster, et al. (1981), Laird and Ware (1982), Mason, et al. (1984), Raudenbush and Bryk (1986), Rubin (1980), and Sternio, et al. (1983). Other numerical approaches to maximum likelihood estimation of covariance components are the iterative generalized least squares (Goldstein, 1986) and the Fisher scoring method (Longford, 1985; de Leeuw and Kreft, 1986).

All these three iterative methods avoid the inversion of large matrices. Thus, they are computationally more feasible than Newton-Raphson, which requires inversion of large matrices at each iteration. S.J. Haberman, one of the discussants of the paper by Dempster et al. (1977), pointed out that the numerical stability and simplicity of implementation of the EM algorithm are in its favor. The Newton-Raphson and scoring algorithms are not especially difficult to implement. However, convergence of the EM algorithm is often slow. In contrast, the Newton-Raphson and scoring method are superior from the point of view of rate of convergence near a maximum since they converge quadratically rather than linearly. However, they do not have the property of always increasing the likelihood, and can in some instances move toward a local minimum. Consequently,

the choice of starting value may be more important under Newton-Raphson and scoring method (Dempster, et al., 1977).

The Logic of EM Estimation

The EM algorithm of Dempster, et al. (1977) provides an iterative method of finding the maximum likelihood variance estimates. The EM algorithm is a very general method for finding maximum likelihood estimates. In the variance estimation situation, the EM algorithm alternates two steps in each iteration. The E ("expectation") step finds the posterior expectation of the sufficient statistics based on the complete data (in our case y, θ) given the observed data (in our case y) and given current estimates of parameters (in our case τ and σ_j^2). The M ("maximizing") step then uses the expected sufficient statistics to produce new ML parameter estimates of variance components. Each step of the EM algorithm increases the likelihood. This sequence of alternate steps guarantees convergence to a local maximum of the likelihood function. If data is normally distributed, the local maximum will also be the absolute maximum since the normal likelihood is unimodal.

One difficulty with EM algorithm is that it may require many iterations to converge (Sternio, 1981). Thus, it is a slow process of maximum likelihood estimation particularly with poor starting values (Mason, et al., 1984). Nonetheless, in favor of the EM algorithm are simplicity in implementation and numerical stability.

The process of the EM algorithm along with the computational details are provided by Dempster, et al. (1981) to a special version of the model considered in this research. They considered the model in which there are no macro predictors (covariates) related to micro parameters, i.e., $W = I$ and in which the Σ_j have the special form: $\Sigma_j = \sigma^2 I_{n_j}$ where σ^2 is equal across all individuals. Sternio (1981) has broaden this appraoch to include estimation of τ and σ^2 in more general cases and provides a unified discussion of theory and computation in such cases. Bryk, Raudenbush, Seltzer, and Congdon (1987) have extended this approach even further to general mixed model in which the full rank assumption of the within-group predictor matrix X_j , and the assumption that micro parameters are random are no longer required. Hence, relaxing these two restrictive assumptions broadens the range of application of the model (Braun, et al., 1983; Rubin, 1983).

To illustrate the logic of EM algorithm, consider the simple conditional univariate HLM model prescribed in equations (3.7) and (3.11). The logic of EM works like this: First, assume that τ and σ_j^2 are known. Equations (3.12) through (3.15) provide posterior means and dispersions of γ and θ_j . Next, suppose that γ and θ_j were known i.e., R_j and U_j had been observed and we want to estimate τ and σ_j^2 . It can be shown easily that the following two equations (3.16) and (3.17) are maximum likelihood estimates for τ and σ_j^2 respectively.

$$\hat{\tau} = K^{-1} \Sigma U_j U_j' \qquad (3.16)$$

$$\hat{\sigma}_j^2 = R_j' R_j / n_j \qquad (3.17)$$

EM algorithm utilizes the dependence of estimators θ_j and γ on knowledge of dispersion matrices and the dependence of ML estimators of these matrices on knowledge of θ_j and γ via an iterative process with the following steps:

(1) Generate reasonable starting values for the unknown variances, σ_j^2 and τ. Perhaps as suggested by Raudenbush (1988), the within-group and between-group residuals from ordinary least squares regression can be used.

(2) These starting values are substituted into equations (3.12) and (3.13) yielding starting values of θ_j^* and γ^*.

(3) To derive new estimates for τ and σ_j^2, substitute the sufficient statistics $\Sigma U_j U_j'$ and $R_j' R_j$ in equations (3.16) and (3.17) by their posterior expectations.

These posterior expectations are derived by Dempster, et al. (1981) and are as follows:

$$\begin{aligned} E\{ (R_j'R_j) \mid Y\} &= E\{(Y_j - X_j\beta_j)' (Y_j - X_j\beta_j) \mid Y\} \\ &= (Y_j - X_j\beta_j^*)'(Y_j - X_j\beta_j^*) + E\{(\beta_j - \beta_j^*)'X_j'X_j(\beta_j - \beta_j^*) \mid Y\} \\ &= (Y_j - X_j\beta_j^*)'(Y_j - X_j\beta_j^*) + tr(X_j'X_j D_{\beta_j}^*) \end{aligned}$$

and

$$E(\Sigma U_j U_j' \mid Y) = \Sigma U_j^* U_j^{*'} + \Sigma\{\lambda_j V_j + W_j S_j W_j'\}$$

(4) The new estimates of σ_j^2 and τ are then used in a repetition of step 2 to yield new value for θ_j^* and γ_j^*.

The process iterates until estimates converge to any degree of accuracy required.

The estimated variance components after one iteration are then the maximum likelihood estimates of the variances, conditional on the values of the structural parameters (i.e., regression coefficients). The proof of convergence to the maximum likelihood estimates is given by Dempster, et al. (1977).

Effects of Having to Estimate Variance Components

Best linear unbiased estimators of the fixed and random effects (i.e., macro and micro parameters respectively) of mixed linear models are available when the true values of the variance components are known. If the true values are replaced by estimated values, the mean squared errors of the estimators of the macro and micro parameters increase in size (Kackar and Harville, 1984). Clearly the magnitude of this increase is unknown to us. Another problem resulting from this situation is that the parametric family distribution of micro and macro parameter estimates will remain unknown. Thus, any statistical inference concerning these parameters, if not impossible, will be inaccurate.

Fortunately, we can use large sample theory to find asymptotic distributions of macro parameter estimates. But finding an analogous sampling distribution for micro parameters is not possible (Dempster, et al., 1981), because we cannot simultaneously maximize the joint likelihood function

of all four parameters (θ , γ , τ and σ_j^2). The data simply will not support the estimation of so many parameters. But the focus of the present research is on the effect of variance estimation on inferences about macro parameters.

Of course, when variance components are unknown, substituting their maximum likelihood estimates $\hat{\tau}$ and $\hat{\sigma}_j^2$ in the definition of $\hat{\Delta}_j$ and then estimating γ^* by replacing Δ_j for $\hat{\Delta}_j$ in equation (3.13) is a natural idea. That is, following empirical Bayes approach of first deriving Bayesian estimates based on known variances and then substituting ML estimates for the unknown variances in the estimation formulas. The resulting empirical Bayes estimator, γ^* is a true maximum likelihood estimator. Therefore, this estimate shares the desirable properties of maximum likelihood estimators. But maximum likelihood estimators rely on large sample theory. According to large sample theory we know that:

1. $$\gamma^* = (\Sigma\, W_j' \hat{\Delta}_j^{-1} W_j)^{-1} \Sigma\, W_j' \hat{\Delta}_j^{-1} \hat{\theta}_j ,$$

is the maximum likelihood estimate of γ^* if $\hat{\Delta}$ is the maximum likelihood of Δ . This is the case since functions of ML estimates are ML estimates of the same functions of the parameters.

2. Under regularity condition the large sample distribution of ML estimates of $\hat{\gamma}^*$ for $K \to \infty$ with n's fixed is as follows

$$\hat{\gamma}^* \sim N\left|\gamma^*, (\Sigma W_j' \Delta_j^{-1} W_j)^{-1}\right|$$

where $(\Sigma W_j' \Delta_j^{-1} W_j)$ is the Cramer Rao lower bound for the covariance matrix of $\hat{\gamma}^*$.

But for n, K $\to \infty$ with n's/N fixed

$$(\Sigma W_j' \Delta_j^{-1} W_j)^{-1} = (\Sigma W_j' \tau^{-1} W_j)^{-1}$$

since

$$\Delta_j^{-1} = (V_j + \tau)^{-1} = \left| \sigma_j^2 (X_j' X_j)^{-1} + \tau \right|^{-1} = \tau^{-1}$$

as

$$\sigma_j^2 (X_j' X_j)^{-1} \to 0$$

thus γ^* is indeed asymptotically efficient. It is clear that we can use the asymptotic distribution of γ^* for confidence interval and hypothesis testing:

$$(\gamma^* - \gamma) \overset{ASY}{\sim} N \left| 0 , (\Sigma W_j' \Delta_j^{-1} W_j)^{-1} \right|$$

or

$$(\gamma_h^* - \gamma_h / S.E. \gamma_h^*) \overset{ASY}{\sim} N (0, 1)$$

where subscript h refers to the elements in the γ vector i.e., $(\gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11})$.

Thus, even though the estimates of the macro parameters are numerically computed, their large sample properties are well defined which facilitates the large sample hypothesis testing and interval estimation.

The EM algorithm yields estimates of dispersion matrices τ and Σ which, in conjunction with γ^* , maximize the marginal density of Y . In other words, EM estimates of τ

and Σ (i.e., maximum likelihood estimates of τ and Σ) when substituted into the equation (3.13) make γ^* a true maximum likelihood estimator. These asymptotic properties of ML estimates are of value only if there is reason to believe that the data are extensive enough that the properties hold. For these properties to hold exactly it is sufficient that the number of groups (K in our case), approaches to infinity (Miller, 1977). However, it would be interesting to observe the behavior of the estimates as K and n (i.e., number of individual within each group) each increase to infinity. This does not imply that K and n be of the same order of magnitude (Miller, 1977). In the present research, the main question we set out to investigate concerns the small sample behavior of the macro estimators (i.e., γ_{00}, γ_{01}, γ_{10} , and γ_{11}). The purposes of this study are three:

(1) To check on the EM algorithm, we can look at the properties of the macro estimators and make sure that the algorithm behaves as expected. That is, the macro estimators are consistent, unbiased, asymptotically efficient, and with known and asymptotic normal distribution. Also it is worthwhile to look at how well the EM algorithm does at estimating the variance components. Again, this concerns the bias, consistency and asymptotic efficiency of thse estimators. A side concern with this algorithm is its rate of convergence under varying combinations of K and n. This question is addressed through examining the total number of iterations prior to convergence to ML estimates.

(2) Investigating the effect of variance estimation on inferences about macro parameters with respect to both robustness and power.

(3) By constructing data sets that differ in the number of K and n, we investigate how different combinations of K and n affect the properties of the macro estimators as well as the inferences about them.

Specific questions of interest concentrate on estimation and hypothesis testing. The key issue in the estimation phase concerns the bias, consistency and efficiency of the macro estimators, γ , and the effect of different combinations of K and n on these properties. For hypothesis testing interest centers on the type I errors and power.

CHAPTER IV

METHOD

The procedures employed in the study to answer the research questions presented in the previous chapter will now be discussed. The chapter begins by presenting standardized two-stage hierarchical linear model. Next, a description of the population parameters and the manner in which they were chosen will be given. In the third section details are presented about the computer routine utilized to generate the data. The fourth section looks at the analysis routines. Finally, the measures of biasedness, consistency, efficiency, type I errors and power will be described.

Standardized Two-Stage Hierarchical Linear Model

This model which is special case of the two-stage HLM model presented in the preceding chapter is adopted for generating data in the present research. The standardized HLM takes the same exact form of equations (3.1) through (3.5) for the unconditional and conditional case but with somewhat different assumptions. That is, the micro and macro predictors both are assumed to be standardized normal variables with mean of zero and variance of one. Clearly, this reduction in the number of unknown parameters simplifies the data generation process.

Within-School Model

$$Y_{ij} = \mu_j + \beta_j X_{ij} + R_{ij},$$

and

$$R_{ij} \sim N (0, 1).$$

Between-School Model (unconditional)

$$\mu_j = \bar{\mu} + U_{0j},$$

$$\beta_j = \bar{\beta} + U_{1j},$$

and

$$U_{0j} \sim N (0, \tau_\mu),$$

$$U_{1j} \sim N (0, \tau_\beta),$$

$$\text{cov} (U_{0j}, U_{1j}) = \tau_{\mu\beta}.$$

Between-School Model (conditional)

$$\mu_j = \gamma_{00} + \gamma_{01} W_j + U_{0j},$$

$$\beta_j = \gamma_{10} + \gamma_{11} W_j + U_{1j},$$

and

$$U_{0j} \sim N (0, \tau_{\mu|W}),$$

$$U_{1j} \sim N (0, \tau_{\beta|W}),$$

$$\text{cov} (U_{0j}, U_{1j}|W) = \tau_{\mu\beta|W}.$$

Further we assume that the micro and macro predictors are each a unit normal random variable, i.e.,

$X_{ij} \sim N(0, 1)$,

$W_j \sim N(0, 1)$,

and that X_{ij}, R_{ij}, U_{0j} and U_{1j} are mutually independent. This implies that $COV(U_{0j}, U_{1j}) = 0$, and that the dispersion matrix τ is diagonal (but we will still investigate estimates of this covariance between macro level errors).

In order to generate data we need to define the following parameters:

1. $c = \tau_\beta / \tau_\mu$, so $\tau_\beta = c\,\tau_\mu$. (4.1)

2. $\bar{\beta}^2 = \bar{\rho}_{xy}^2 \left(\bar{\sigma}_y^2 / \bar{\sigma}_x^2 \right)$

where

$\bar{\beta}$ is pooled within group slope;

$\bar{\rho}_{xy}$ is pooled within group correlation coefficient;

$\bar{\sigma}_y^2$ is pooled within-group (unconditional) variance in Y ; and

$\bar{\sigma}_x^2$ is pooled within variance in X .

But

$\bar{\sigma}_x^2 = 1$,

and

$$\bar{\sigma}_y^2 = \tau_\beta + \bar{\beta}^2 + \text{var}(R) \qquad (4.2)$$
$$= c\tau_\mu + \bar{\beta}^2 + 1,$$

so that

$$\bar{\beta}^2 = (\bar{\rho}^2_{xy} / (1 - \bar{\rho}^2_{xy}))(c\tau_\mu + 1). \qquad (4.3)$$

$$3.\ d = \tau_\mu/(\tau_\mu + \bar{\sigma}^2_y), \qquad (4.4)$$

where d is the intraclass correlation of Y.

By substituting expression (4.3) into expression (4.2) we will get:

$$\bar{\sigma}^2_y = c\tau_\mu + (\bar{\rho}^2_{xy}/(1 - \bar{\rho}_{xy}))(c\tau_\mu + 1) + 1,$$

and by substituting $\bar{\sigma}^2_y$ into (4.4) and solving for τ_μ we will have:

$$\tau_\mu = d/((1-d) (1 - \bar{\rho}^2_{xy}) - cd). \qquad (4.5)$$

This expression implies that the larger the intraclass correlation the larger the parameter variance τ_μ , and that the larger the pooled within-group correlation the smaller the τ_μ . That is, τ_μ is directly related to d , but inversely to $\bar{\rho}^2_{xy}$.

But τ_μ and τ_β are both positive quantities, therefore c is constrained to be in the following range of values

$$0 < c < ((1 - d)/d)(1 - \bar{\rho}^2_{xy}). \qquad (4.6)$$

Also the conditional parameter variance in intercept and slope are:

$$\tau_{\mu|W} = \tau_\mu(1 - \rho^2_{\mu W}), \qquad (4.7)$$

and

$$\tau_{\beta|W} = \tau_{\beta}(1 - \rho^2_{\beta W}), \quad \text{respectively.} \qquad (4.8)$$

The above specification of the standardized hierarchical linear model reduces to five parameters. These parameters are c , d , ρ_{xy} , $\rho_{\mu W}$ and $\rho_{\beta W}$; if predetermined, one can generate a large number of samples under these known population parameters and investigate the properties of resulting statistics (i.e., point estimates of and their standard errors) by observing their sampling distributions.

Parameters of the Study

In order to investigate the small sample properties of the macro parameters (i.e., γ_{00} , γ_{01} , γ_{10} and γ_{11} , two more parameters need to be added to the list of five model parameters previously mentioned. These two parameters are: number of groups, K and group size, n . This adds up the number of parameters considered in the present study to total of seven (K, n, d, ρ_{xy}, $\rho_{\mu W}$, $\rho_{\beta W}$, and c).

The first three parameters (K, n and d) are specially of great concern in the present study because of their significant implications in sampling and design of a study. In a two-stage random sampling (or two-stage cluster sampling using sampling design terminology), the coefficient of intraclass correlation (d) measures the homogeneity of the elements within clusters. For a fixed total sample size of N = nK, the larger the intraclass correlation, the larger

the number of groups (K) and the smaller the number of individuals within groups need to be sampled for optimum efficiency in design given fixed cost. In contrast, the smaller the d , the fewer the number of groups and the larger the number of individuals within groups the better the precision (Kish, 1983).

However, to consider asymptotic properties of macro estimators it is sufficient that only K converges to infinity (Miller, 1977). Accordingly, we may occasionally define the population solely in terms of levels of K, on other occasions in terms of varying combinations of K and n, and still on some occasions redefine it in terms of all three parameters. Now the values assigned to each of these parameters will be given.

(1) Number of Groups, K. Small to moderate to large groups with K = 10, 30, 60 and 150 are simulated in this study.

(2) Group Size, n. Situations with n = 5, 25, 60 and 150 in each group are simulated.

In deciding the values of K and n, the main concern was to select those values that provide us with reasonable ground to investigate the small sample properties of the macro parameters of interest. The other concern was to have a reasonable coverage of those combinations of n and K which occur in real research situations.

These realistic situations include; 1) the study of growth model (Bock, 1983; Goldstein, 1986; Laird and Ware, 1982; and Sternio, et al., 1983) in which K is small and n

ranges from small, moderate to large; 2) school effects research (Aitken and Longford, 1986; Raudenbush and Bryk, 1986), where K is moderate or large and n is either small or moderate; and 3) sociological and contextual research (Mason, et al., 1984; and Wong and Mason, 1985) in which K is small to moderate and n is large.

3) Intraclass Correlation of Y, d . Two values of d =.10 and .25 are considered. These two values appear to be of reasonable magnitude based on the following grounds. The intraclass correlation of Y may be large if τ_μ is large compared with $\bar{\sigma}^2_y$, and zero only when $\tau_\mu = 0$, that is when there is no variation in outcome variable among schools in the population of schools, which will rarely happen in practice (see expression 4.4) . But as a general rule, intraclass correlations in educational research are small positive values, mostly under .15 (Kish, 1983).

The range of the values are chosen to reflect the values often obtained from educational field research. In the school effect research conducted by Raudenbush and Bryk (1986) the actual value of intraclass correlation was .177. The mid point of the values considered in this study is .175.

4), 5) and 6) Correlation Coefficients of ρ_{xy} , $\rho_{\mu W}$, and $\rho_{\beta W}$. For each of these correlation coefficients two values are considered. These values which are considered to be of moderate and almost high magnitude (considering educational field research data), are .25 and .75.

7) Ratio of τ_β to τ_μ , c . Two values of c =.10 and .50 are considered in this study. Both fall within the range of permisable values for c given by expression (4.6). Two interrelated factors have affected the selection of these two values: First, as a general rule, regression coefficients have considerably greater sampling variability than sample means(Burstein and Miller, 1980; Wiley, 1970). Mathematically, the total variability in intercepts and slopes can be decomposed into two parts; parameter variance and sampling variance. Logically it follows that the parameter variance in intercept is of larger magnitude than that of slope. Second, in many applications one would expect that much of the observed variation in slopes to be sampling variation. For example, in the school effect research conducted by Raudenbush and Bryk (1986) which utilized a sample of 10231 students in 176 schools, student samples per school ranged from 10 to 70, and samples less than 45 were rare, and the value of c was equal to .10. Consequently, this value is chosen in this study to act as a baseline, and will be compared to a less realistic but certainly not impossible larger value of c i.e., .50.

Design of the Study

Considering the number of factors (total of seven) and number of levels in each factor (K and n each have 4 levels, and the remaining 5 factors each have 2 levels), if we were to include all factor combinations in our study, we would have a $(2^5 \times 4^2)$ design matrix with a total of 512 possible

cells, which is unmanagable given the large cost of implementing the EM algorithm.

As a practical alternative, this study adopted a fractional factorial design by which only a fraction of factor combinations of a complete fractional design will be considered. Specifically, this study has adopted a "one-half" randomized block fractional factorial (RBFF).

Kirk (1968:386-87) made the following comment concerning fractional factorial designs: "the use of a fractional design can lead to a sizeable reduction in the number of treatment combinations that must be included in a study. This is accomplished by confounding main effects with higher order interactions..... however, if certain information concerning the outcome of the experiment is of negligible interest, an experimenter can employ confounding so as to sacrifice only this information."

As a result of treatment-interaction confounding, considerable ambiguity may exist in interpreting the results of such experiments. This is the case since every sums of squares can be given two or more designations referred to as "aliases". To minimize this ambiguity, careful attention must be given to the alias pattern of a proposed design. Treatments are customarily aliased with next to highest-order interactions which can be assumed to equal zero. This is accomplished by using the highest-order interaction as the "defining contrast" which is used to divide the treatment combination into two blocks. The higher order interac-

tions are then pooled to form a residual error term. "If these pooled interactions are insignificant, a complete factorial design would have been a better design choice for the data than the fractional factorial design. On the other hand, if some of the interactions are significant, the present analysis (i.e., fractional factorial design) offers the advantage of a larger number of degrees of freedom for experimental error and a within-all error term" (Kirk, 1968: 394).

Designs with mixed treatments (or factors), i.e., having unequal number of levels, present special problems with respect to layout and analysis (see Kempthorne, 1952: 419). But this is the case in the present study which contains mixed treatments of the form $2^5 \times 4^2$ design. As a reasonable alternative this study, adopted a RBFF- 2^5 design for the five factors with two levels (i.e., c , d , ρ_{xy} , $\rho_{\mu W}$, and $\rho_{\beta W}$), and to compensate for the two remaining factors, K and n each with four levels, every two blocks of the design layout of RBFF- 2^5 was crossed with different level combinations of K and n.

Next, steps involved for laying out one-half replication of a type RBFF- 2^5 fractional factorial will be discussed. (1) Choose a defining contrast. Following Kirk's guideline the highest order interaction (i.e., five order interaction) is chosen in this study as the defining contrast. The 32 treatment combinations ($2^5 = 32$) of a complete factional design can be reduced to one-half of that by the use of the

defining contrast.

(2) Confound an interaction with between-block variation. The interaction which serves as the confounding interaction must be insignificant and also different from the defining contrast. For this purpose the interaction between $\rho_{\mu W}$ and $\rho_{\beta W}$ is chosen which is thought to be insignificant. For confounding an interaction with blocks see Kirk (1968, Chapter 9). As a result of this process, the 16 treatment combinations are assigned to two blocks of eight combinations each.

For simplicity the five factors are assigned the following notations:

$A = \rho_{\mu W}$

$B = \rho_{\beta W}$

$C = \rho_{xy}$

$D = d$

$E = c$

If (ABCDE) is used as the defining contrast and (AB) as confounding interaction, the design shown in Table 4.1 will be obtained. Levels of each factor are denoted as zero and one. Where zero corresponds to the low value and one to the high value.

All treatments and interactions except AB (the confounding interaction), its alias CDE, and the defining contrast ABCDE are within-block effects. All main effects are aliased with four-factor interactions. The alias

Table 4-1

Layout of Type RBFF - 2^5 Design in Two Blocks
Treatment Combination

	abcde	abcde	abcde	abcde	abcde	abcde	abcde	abcde
Block 0	00000	00110	00101	00011	11000	11110	11101	11011
Block 1	01111	01100	01010	01001	10100	10010	10001	10111

Table 4-2

Alias Pattern for Type RBFF - 2^5 Design

Source	Blocks (AB)	A	B	C	D	E	AC	AD	BC	BD	CD	ABC	ABD	ACD	BCD
Alias	CDE	BCDE	ACDE	ABDE	ABCE	ABCD	BDE	BCE	ADE	ACE	ABE	DE	CE	BE	AE

pattern for this design appear in Table 4.2.

A careful examination of the alias pattern in Table 4.2 reveals an interesting feature of this one-half fractional factorial design. The incomplete five-treatment design contains all of the treatment combinations of a complete four-treatment design. This implies that the computational procedures for a one-half replication of a 2^5 design are identical to those for a complete replication of a 2^4 design. That is, by ignoring one of the treatments, the analysis of an incomplete-design can be carried out as if all the treatment combinations were included in the experiment. The choice of which treatment to ignore is arbitrary (Kirk, 1968).

As mentioned earlier, different combinations of K and n will be crossed with the blocks contained in the RBFF- 2^5 design. There are a total of 4^2 different combinations of K and n, each referred to as a "trial" for convenience. Within each trial the first K by n level combination will be crossed with "block 0" of the RBFF design and the second K by n level combination will be crossed with "block 1". Table 4.3 contains all different combinations of K and n, their designated block, along with the trial number.

By using a one-half RBFF- 2^5 design and by crossing this design with particular combination of K and n a total of 128 (16 cells x 8 trials = 128) treatment combinations will result. Notice that although levels of K are crossed with both blocks of the design matrix, levels of n are not.

Table 4-3

Varying Combination of K and n
for Blocks 0 and 1 of RBFF - 2^5 Design

	Trial							
	1	2	3	4	5	6	7	8
Block 0	k=10 n=5	k=10 n=60	k=30 n=5	k=30 n=60	k=60 n=5	k=60 n=60	k=150 n=5	k=150 n=60
Block 1	k=10 n=25	k=10 n=150	k=30 n=25	k=30 n=150	k=60 n=25	k=60 n=150	k=150 n=25	k=150 n=150

Table 4-4

Actual Values of Macro Parameters

Block 0	γ_{00}	0	0	0	0	0	0	0	0
	γ_{01}	.08658	.22704	.13484	.16440	.25974	.68111	.40452	.49320
	γ_{10}	.25974	1.17973	1.21356	.28473	.25974	1.17973	1.21356	.28475
	γ_{11}	.02738	.07180	.09535	.11625	.08214	.21539	.28604	.34874
Block 1	γ_{00}	0	0	0	0	0	0	0	0
	γ_{01}	.27735	.12762	.15179	.08874	.38276	.45538	.26621	.83205
	γ_{10}	1.44115	1.14857	.26292	.26621	1.14857	.26292	.26621	1.44115
	γ_{11}	.58835	.12107	.14400	.18824	.04036	.04800	.06275	.19612

Specifically samples of size 5 and 60 occur only in "block 0" and 25 and 150 in "block 1". Thus, each trial consists of either the two lowest levels of n, 5 and 25 (call it n'), or the two highest levels 60 and 150 (call it n").

As a result of this design with sixteen varying combinations of factors A, B, C, D, and E, we obtain sixteen different parameter values for γ_{01} and γ_{11} as shown in Table 4-4. With regard to γ_{10} the total number of parameter values reduces to half of this size since γ_{10} is defined only in terms of C, D, and E. Thus, it is not affected by the high and low values of A and B. Irrespective of the design γ_{00} is pre-fixed at zero.

Description of the Generation Routine

In generating data the present study make use of five sufficient statistics. Generally speaking, sufficient statistics are useful in that they reduce the number of observations, say from n to r statistics (where $r \leqslant n$). This is because these r statistics contain all the "information" about θ (i.e., parameters of the study) that the n observations contain (Graybill, 1976). If r is appreciably less than n , as it is in the present study (i.e., r = 1/n), then the very fact that we have to consider only r , rather than n simplifies our data generation routine.

The five sufficient statistics are ΣX , ΣX^2 , ΣR , ΣR^2 , and ΣXR . Assume that,

$X \overset{iid}{\sim} N(0, 1)$,

$R \overset{iid}{\sim} N(0, 1)$,

and

$\rho_{XR} = 0$

(i.e., the population correlation coefficient is zero).

The generation procedure is composed of the following steps:

(1) Generate

$\Sigma X_j \sim N(0, n)$.

(2) First generate $\Sigma(X_j - \bar{X}.)^2 \sim \chi^2_{(n-1)}$ and then compute

$$\Sigma X_j^2 = \Sigma(X_j - \bar{X}.)^2 + (\Sigma X_j)^2/n .$$

(3) Generate

$\Sigma R_j \sim N(0, n)$.

(4) First generate $\Sigma(R_j - \bar{R}.)^2 \sim \chi^2_{(n-1)}$ and then compute

$$\Sigma R_j^2 = \Sigma(R_j - \bar{R}.)^2 + (\Sigma R_j)^2/n.$$

(5) To generate ΣXR first generate t with (n-2) degrees of freedom, then compute

$$r = t / (t + n - 2)^{1/2},$$

and finally compute

$$\Sigma X_j R_j = (n - 1)\, r\, S_x\, S_R + \Sigma X_j R_j / n \ .$$

(note: $t = Z/(\chi^2_{(n-2)}/(n-2))^{1/2}$)

After completing the steps involved in generating the sufficient statistics, we can actually compute ΣY, ΣY^2 and ΣXY. But before doing so we need to generate three more random variables which are contained in the conditional between-group model. These random variables, which are part of the expression for μ_j and β_j and thus part of Y_{ij}, are W_j, U_{0j} and U_{1j}. Also we need to assign values to the four macro parameters of interest. Assignment of values to the slopes, γ_{01} and γ_{11} are accomplished through the following expressions:

$$\gamma_{01} = \rho_{\mu W} \sqrt{\tau_\mu}$$

$$\gamma_{11} = \rho_{\beta W} \sqrt{\tau_\beta}$$

γ_{00} is assumed equal to zero, and γ_{10} is assumed equal to $\bar{\beta}$. Now proceed with the steps in generation routine.

(6) Generate

$$W_j \sim N(0, 1).$$

(7) Generate

$$U_{0j} \sim N(0, \tau_{\mu|W}).$$

(8) Generate

$$U_{1j} \sim N(0, \tau_{\beta|W}).$$

Also notice that prior to generating U_{0j} and U_{1j} we need to assign values to the five model parameters; i.e., c , d , ρ_{xy} , $\rho_{\mu W}$ and $\rho_{\beta W}$. The final step in generation routine is to compute ΣY , ΣY^2 and ΣXY .

(9) Compute

$$\Sigma Y = \Sigma(\mu_j + \beta_j X_{ij} + R_{ij})$$

$$= n\mu_j + \beta_j \Sigma X_{ij} + \Sigma R_{ij}.$$

Compute

$$\Sigma Y^2 = \Sigma(\mu_j + \beta_j X_{ij} + R_{ij})^2$$

$$= n\mu_j^2 + \beta_j^2 \Sigma X_{ij}^2 + \Sigma R_{ij}^2 + 2 \mu_j \beta_j \Sigma X_{ij} +$$

$$2 \mu_j \Sigma R_{ij} + 2\beta_j \Sigma X_{ij} R_{ij}.$$

Compute

$$\Sigma XY = \mu_j \Sigma X_{ij} + \beta_j \Sigma X_{ij}^2 + \Sigma X_{ij} R_{ij}.$$

The generation program, completes each of the nine steps as one observation is formed. The sample size chosen is five so five such vectors Y compromise one sample. Thus, beginning with the first "trial", values of 10 and 5 will be assigned to K and n respectively of "block 0" in the RBFF-2[5] design, and similarly values of 10 and 25 to K and n of "block 1". Then, starting with the cell one of the design layout, first the remaining five parameters will be assigned

values (according to zero and ones), and then the nine data generation steps will be completed and repeated for five replications. This process will be repeated for each and every 16 cells as one trial is completed. A total of 80 (16 cells x 5 observations) data points will be generated upon the completion of this trial. Next, we move to the second trial, assign values to K and n and repeat the same cycle as in the first trial. This process continues until all eight trials are completed and a total of 640 (80 sample points in each trial x 8 trials) sample points is generated. In other words 128 (16 cells x 8 trials) distinct samples each containing five replications wil be generated.

Along with the generation of sample points, the generalization program will compute two indices of dispersion in macro parameters. These indices are the mean squares within and the Cramer Rao lower bound which is

$$- \left| E \left(\partial^2 \log L / \partial \gamma' \gamma \right) \right|^{-1},$$

and equal to the asymptotic dispersion of γ^* i.e., $(\Sigma W_j' \Delta_j^{-1} W_j)^{-1}$.

The first analysis routine (i.e., HLM program) accepts both raw data and summary statistics of the sample means and sample covariance matrix. Considering the efficiency of summary statistics, for each sample, the mean and the covariance matrix is computed to be used as input in the analysis phase.

Monte Carlo Techniques

As recognized by Hammersley and Handscomb (1964), a Monte Carlo method is a general technique with different areas of application, for solving a model by using a random (or pseudo-random) numbers. One application is the generation of sampling distribution. Through repeated sampling under known population parameters, one can investigate the properties of estimators by observing their empirical (sampling) distribution.

The present study is a Monte Carlo study aimed at generating sampling distributions of the macro estimators. These empirical distributions are then compared to the nominal distribution (in this case the normal distribution) obtained under asymptotic theory (i.e., when K and n converge to infinity).

A Fortran program is used to generate a total of 640 sample points; five observations for every 128 experimental conditions.

Random Number Generation

The use of random number is considered to be an integral part of a Monte Carlo study. Random numbers are of two types: purely random numbers and pseudo-random numbers. However, for a computer based Monte Carlo study the purely random numbers are inefficient compared with pseudo random numbers. There are two advantages in using pseudo-random numbers: (1) the computer itself can generate sequence of

numbers by applying an algorithm, and (2) the same sequence of numbers can be reproduced exactly for the future use.

Pseudo-random numbers are generated sequentially from a completely specified algebraic formula. At best which at best they behave as if they are random (i.e., uniformly distributed and mutually independent). These algebraic formulas are devised in such a way to resist any significant deviation from randomness. However, there are many statistical tests that can be used to determine if this is the case. Typically run tests, serial tests, and various Chi-square tests for independence are applied to relatively short sections of the pseudo-random sequence. See Knuth (1969) for discussions on many of these tests.

Two subroutines, GGNML and GGCHS from the International Mathematical and Statistical Library (IMSL) were used to obtain a sequence of pseudo-random normal (R), distributed N (0,1), and Chi-square random deviates with n degrees of freedom respectively. Once the procedure is started by an initial number, called the seed, each new seed number will be determined from the previous one.

Random normal (0, n) deviates can be obtained by transforming GGNML output according to $Y(I) = R(I) \times n^{1/2}$, for I in (1, 2, ..., K). This transformation was done in steps (1) and (3) of the generation routine. In steps (7) and (8) a similar transformation was performed of the form

$$Y(I) = R(I) \times V^{1/2}$$

where V represents $\tau_{\mu W}$ or $\tau_{\beta W}$ whichever the case may be.

Analysis Routine

Output from the generation program consists of summary statistics for each sample. This serves as input to the first analysis routine.

From the first routine, HLM (Bryk, et al., 1987), we obtain a vector of the empirical Bayes estimates of the macro parameters, γ^* (as in equation 3.13), the empirical Bayes estimates of their dispersion matrix, D^*_γ (as in equation 3.15), estimates of parameter variances τ_μ and τ_β, estimate of σ^2, and number of iterations . These estimates are numerically computed via EM algorithm. The convergence criterion for the log likelihood function was set at .0001 with the maximum number of iterations allowed fixed at 500. The empirical Bayes estimates of the macro parameters and their dispersion, parameter variances and σ^2, yet serve as input to the second analysis routine.

The analysis routine computes: (1) the required summary statistics for the estimation phase, (2) the proportion of times the values of each test statistics exceeded its critical values for a given nominal significance level under true null hypothesis, (3) the noncentrality parameter (ncp) as is defined in the last section of this chapter, and (4) tabulates population effect size (γ in our case) against ncp as a way of demonstrating power functions.

Checking the EM Algorithm

As a check on the algorithm, first we might wish to examine the properties of the numerically computed estimates of the macro parameters, γ . The ML estimators are functions of every sufficient statistic and are consistent and asymptotically normal and efficient. Additionally, given normal data, ML estimates of regression coefficients are unbiassed.

Key issues in estimation concentrate on bias and efficiency of an estimator. An estimator is unbiased if its expected value is equal to the population value of the parameter. In other words, if an estimator is unbiased, the estimated value minus its parameter value should have zero expectation, i.e.,

$$E (\hat{\gamma} - \gamma) = 0$$

where

$$\gamma = \left| \gamma_{00}, \gamma_{01}, \gamma_{10}, \gamma_{11} \right|$$

Thus, by deviating parameter estimate from the known population value and averaging over the entire sample, one can determine the degree of bias, if any, present in the estimation procedure.

An estimator is said to be relatively efficient if it has the smallest standard error term among the set of unbiased estimators. Three estimates of the variance are computed. The first is the variance of the macro estimators estimated by HLM via EM algorithm:

$$\text{var}\left(\hat{\gamma}^*\right) = \text{Diag}\left(\Sigma W_j' \hat{\Delta}_j^{-1} W_j\right)^{-1}$$

where

$$\hat{\Delta}_j = \hat{\tau} + \hat{V}_j$$

To the extent that variance components $\hat{\tau}$ and $\hat{V}_j$ are misestimated due to small sample problem, the estimated variance of macro estimators will be in error. The second estimate of variance is the mean squares within (MSW):

$$\text{MSW} = \sum_{i=1}^{5} \left(\hat{\gamma}^* - \bar{\hat{\gamma}}^*\right)^2/5$$

where

$$\bar{\hat{\gamma}}^* = \sum_{i=1}^{5} \hat{\gamma}^*/5$$

Last measure is the average squared bias or mean square error (MSE):

$$\text{MSE} = \sum_{i=1}^{5} \left(\hat{\gamma}^* - \gamma\right)^2/5$$

These last two measures of variance are similar except that MSE takes advantage of the fact that the population value (γ) is known.

Since the maximum likelihood estimates are asymptotically normally distributed, it is of interest to discover whether they are asymptotically efficient in the sense of attaining the Cramer-Rao lower bound for the covariance matrix. This minimum variance bound is the inverse of the Fisher information matrix and is equal to

the asymptotic dispersion of γ^*, i.e., $(\Sigma W_j' \Delta_j^{-1} W_j)^{-1}$. Consequently, all three measures of variance are averaged over the entire sample and then compared with the asymptotic variance of the macro parameters, i.e., diagonal elements of matrix $(\Sigma W_j' \Delta_j^{-1} W_j)^{-1}$. Computational formulae for these various measures of dispersion are:

1) $\mathbf{VAR} = \sum_{j=1}^{K} \sum_{i=1}^{n} \text{var} (\hat{\gamma}^*) / Kn$

where K is number of groups, and n is group size.

Average estimate for the variance of the macro estimators estimated by HLM via EM algorithm from each sample.

2) $\mathbf{MSW} = \sum_{j=1}^{K} \sum_{i=1}^{n} (\hat{\gamma}^* - \bar{\hat{\gamma}}^*)^2 / Kn$

Average estimate for the variance of the macro estimators which is based on the squared difference between the estimates and the mean estimate.

3) $\mathbf{MSE} = \sum_{j=1}^{K} \sum_{i=1}^{n} (\hat{\gamma}^* - \gamma^*)^2 / Kn$

Average estimate for the variance of the macro estimators which is based on the squared difference between the estimators and the population parameters.

4) $\mathbf{CRLB} = \text{Diag} (\Sigma W_j' \Delta_j^{-1} W_j)^{-1}$

Average of values for the minimum variance bound of the macro estimators.

ML estimators of macro parameters have another desirable property, their asymptotic sampling distribution is known and normal. That is,

$$(\gamma^* - \gamma) \overset{ASY}{\sim} N \left| 0, (\Sigma W_j' \Delta_j^{-1} W_j)^{-1} \right|$$

or equivalently

$$((\gamma_h^* - \gamma_h)/S.E. (\gamma_h^*)) \overset{ASY}{\sim} N (0, 1)$$

where subscript h refers to the elements in the γ vector, i.e., (γ_{00}, γ_{01}, γ_{10}, γ_{11}).

One way to assess this property is through the use of normal probability plots. Similarly, we can determine the degree of bias in variance components produced by EM algorithm. But the estimates of variance components are unstable. The size of the sampling variance of these estimates depends on the size of the parameter variances they estimate, i.e., the larger the parameter variances, the larger the sampling variance of the statistics. To stabilize these estimates a logarithmic transformation is performed on each of these variance components. This then is followed by making correction for bias. Formulas for τ_μ , τ_β and σ^2 are given below:

$\log \hat{\tau}_{\mu|W} - \log \tau_{\mu|W} + 1/K$,

$\log \hat{\tau}_{\beta|W} - \log \tau_{\beta|W} + 1/K$,

$\log \hat{\sigma}^2 - \log \sigma^2 + 1/nK$,

(note: $\log \sigma^2 = \log 1 = 0$).

where 1/K, 1/K and 1/nK are Cramer Rao lower bound for $\log \hat{\tau}_{\mu|W}$, $\log \hat{\tau}_{\beta|W}$, and $\log \hat{\sigma}^2$ respectively, and are used for bias corrections (Pitman, 1938). (note: $E(\log(S^2)) \neq \log \sigma^2$ but $E(\log(S^2) + 1/V) = \log \sigma^2$ where V is the correction for bias in S^2). The efficiency of these variance components will be examined by plotting the log of the squared error estimates in $\tau_{\mu|W}$, $\tau_{\beta|W}$, and σ^2 against their respective asymptotic variance, 2/K , 2/K, and 2/nK (see Bartlett and Kendall, 1946, for derivation of these asymptotic variances). As a last check we look at the EM convergence rate under varying combinations of K and n.

Type I Error Rate and Power

There are two ways to commit an error when making an inference: (1) rejecting a null hypothesis when is true (type I error), and (2) not rejecting a null hypothesis when is false (type II error). Where

α = probability of type I error,

β = probability of type II error,

and $1 - \beta$ = power

An experimenter wants to avoid errors and select a statistical procedure which is powerful enough to detect an "experimental effect" if it exists and in which the level of significance (α) is accurate, i.e., neither inflated, nor conservative. One empirical question is what effect does the estimated variance components have on type I error and power.

Three specified significance levels .01, .05 and .10 are considered in this study . For a given nominal alpha, (100 α)% of the values in a test statistic's distribution will exceed the appropriate critical value under a true null ($H_0 : \gamma = \gamma^*$) where $H_a : \gamma = 0$) with known variance components. Actual significance level relates to the proportion of the values in a test statistic's distribution that exceed the appropriate critical value under true null and estimated variance components. Hence, an empirical estimate of the probability of type I error (i.e., actual significance level under unknown variance components) is determined by counting the frequency with which the test statistics ($Z = (\gamma^* - \gamma)/S.E. (\gamma^*)$) in each replication exceeds the corresponding critical value, and the dividing by the total number of replications.

Nominal power relates to the proportion of the values in a test statistic's distribution that exceed the appropriate critical value under a true alternative ($H_a : \gamma = \gamma^*$ where $H_0 : \gamma = 0$) and known variance components. Notice that the null and the alternative hypotheses are the same as the ones under robustness but have switched their position. An empirical estimate of power (i.e., actual power when variance components are estimated) is determined by counting the frequency with which the observed test statistics ($Z_{OBS} = (\gamma^*-\gamma)/S.E.(\gamma^*)$) in each replication exceeds the corresponding critical value, and then dividing by the total number of replications. This count is made at all three

nominal significance levels.

Power is a function of the discrepancy between central and noncentral distribution for a test statistics. In this study actual noncentrality parameters (ncp_a) is defined as the expected value of the observed test statistics i.e.,

$$ncp = E\ (\ Z_{OBS}\ ) = E\left|\ \gamma^*/\ S.E.\ (\ \gamma^*\ )\right|$$

where S.E. (γ^*) is the standard error of the macro estimators estimated by HLM via EM algorithm,

and

$$Z_{OBS} \sim N\ (ncp,\ 1)$$

Actual power under noncentral distribution and unknown variance components is simply equal to the probability of Z_{OBS} exceeding the corresponding critical values:

$$\text{Actual power} = P\ (\left|\ Z_{OBS}\right| > C.V.\ (\alpha/2)\ )$$

where

$$Z_{OBS} - ncp_a = Z, \quad \text{and} \quad Z \sim N\ (\ 0,1\ )$$

Empirical estimates of actual power are then compared to the nominal power in which the nominal noncentrality parameter (ncp_n) is defined as:

$$ncp_n = E\ (\ \gamma^*/\sigma\ (\ \gamma^*) = \gamma/\sigma_\gamma$$

where $\sigma_{(\gamma^*)}$ is square root of the asymptotic dispersion of γ^* (i.e., the Cramer Rao minimum variance bound).

CHAPTER V

RESULTS

The results of the study are presented in this chapter in three sections. The first section is a check on the EM algorithm examining the properties of the maximum likelihood estimates of macro parameters and variance components. The second section will address robustness and power issues and the implication of variance estimation on inferences about macro parameters. The last section presents the rate of convergence of the EM algorithm under varying combinations of K and n used in this study.

Results for Estimation Phase

The objective for this phase of the study is to check the EM algorithm with respect to macro parameters and variance components (the vector notation $\underset{\sim}{\gamma}$ and $\hat{\underset{\sim}{\gamma}}$ are used throughout this section to refer to the macro parameters γ_{00}, γ_{01}, γ_{10} and γ_{11}, and their estimates $\hat{\gamma}_{00}$, $\hat{\gamma}_{01}$, $\hat{\gamma}_{10}$ and $\hat{\gamma}_{11}$ respectively). The question could be phrased: Does the algorithm behave as expected ? That is, are the macro estimators (1) unbiased, i.e., $E(\hat{\underset{\sim}{\gamma}}-\underset{\sim}{\gamma}) = 0$; (2) asymptotically efficient; and (3) with known and asymptotic normal distribution? Operationally this question could be phrased:

Does the estimation get better as a function of K and n ? That is, are the macro parameters consistent, less biased, and more efficient ? Similar questions will be addressed with regard to the variance components, $\hat{\tau}_{\mu|W}$, $\hat{\tau}_{\beta|W}$ and $\hat{\sigma}^2$.

Are the Macro Parameters Asymptotically Unbiased and Consistent ? The error estimates in macro parameters are calculated by subtracting the estimated values $\hat{\gamma}$ from their corresponding parameter values γ . These values are then averaged over the entire sample, 640 sample points. The expected errors of all four macro parameters and their 95 percent confidence intervals are shown in Table (5-1) which suggest that the maximum likelihood estimates of the macro parameters are unbiased.

Table 5-1

Expected Errors of Estimate in the Macro Parameters*

γ_{00}	γ_{01}	γ_{10}	γ_{11}
.002178	.003224	-.001964	.005195
(-.0078, .0118)	(-.0088, .0148)	(-.0098, .0058)	(-.0028, .0128)**

*From 640 replications
**95% confidence intervals

To assess the differential effects of K, n and, d on the error of estimates and to examine more explicitly the differences among levels of each factor, Tables 5-2 through

5-5 give error of estimates in $\hat{\gamma}$ reflecting these three factors. In all four tables the same patterns emerged. Since this was consistent across all four macro parameters, only the results for the $\hat{\gamma}_{00}$ will be discussed. The first eight rows relate to the averaged within-cell error of estimates. Generally these values considering the small number of replications (five) tend to be small. In the lower part of the tables absolute errors of estimates are summed within: 1) levels of d ; 2) levels of n (n' vs. n"); and 3) levels of K. Within each level of n, error increased as d increased. For example, with K = 10 under n' the absolute error of estimates in $\hat{\gamma}_{00}$ (Table 5-2) went from .517 under low intraclass correlation to 1.013 with a high degree of d (d = .25). With K = 30, 60 and 150 under n' absolute error increased from .325, .112 and .145 to .558, .276 and .172,respectively.

This upward trend was remarkably consistent among all macro parameters. The only difference among the four parameters was one of magnitude. With γ_{00} and γ_{01} , error of estimates tended to be slightly higher than for the γ_{10} and γ_{11} . The reason for this difference in reduction in errors for slopes and intercepts seems to be due to the assumed ratio of the parameter variance in slope to that of intercept used to generate data. This ratio is prefixed to be either 1/10 or 1/2 (i.e., c = .10, .50).

Having shown that error of estimates respond differently to the levels of d , the following is an attempt to

Table 5-2

Error Estimates in $\hat{\gamma}_{00}$*

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
.015	.164	.027	.146	.028	.059	.029	.021	.026	-.115	.026	.010	-.018	-.046	.000	.032
-.014	.191	.075	-.053	.108	-.013	.005	-.017	.015	.007	-.004	-.034	.057	-.046	-.044	-.022
.129	.033	-.016	-.143	.068	-.061	-.024	.021	-.003	.008	-.014	-.013	.002	.020	.010	.024
.170	-.179	-.014	.009	-.006	.025	.038	-.051	.011	.021	.025	.021	.002	-.006	.001	-.008
.087	-.315	-.082	.203	.069	-.237	-.092	.034	.004	.041	-.023	-.131	-.023	.013	-.033	.037
-.007	-.067	.043	-.153	.025	.026	.010	-.115	.026	.010	.008	.003	.028	.026	.004	.001
-.084	.049	.004	-.007	.005	-.005	.037	.023	.016	.019	.017	-.000	.004	-.000	-.004	.014
.011	-.015	.001	-.013	.016	-.132	.024	.105	.011	.055	-.013	-.013	.011	-.015	.003	.015
.517	1.013	.262	.727	.325	.558	.259	.387	.112	.276	.130	.225	.145	.172	.099	.153**
1.53		.989		.883		.646		.388		.355		.317		.252***	
2.519				1.529				.743				.569****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

Table 5-3

Error Estimates in
Y_{01}*

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
.020	-.148	-.049	.169	-.022	-.136	.051	-.041	.017	-.091	-.024	.174	.013	-.044	.009	-.029
-.011	.249	.026	.051	.055	.023	-.090	-.096	-.044	.023	.066	-.017	.020	.019	-.007	-.045
.002	.129	-.022	.028	.048	-.072	-.010	.010	.055	-.024	-.007	-.065	.003	.030	.020	.013
.139	.182	-.014	-.007	.027	.005	-.011	-.026	.019	-.055	.011	-.020	-.023	-.045	.009	-.006
-.007	.305	.094	-.105	-.030	-.062	-.011	-.058	.056	-.020	-.004	-.128	.010	-.030	-.002	.008
-.054	.097	.051	.026	.003	-.029	.064	-.064	.003	-.026	-.009	.032	.009	.011	-.022	.022
.124	-.031	.126	-.097	-.040	.012	-.059	.054	.002	-.014	.003	.010	-.010	.007	-.017	.018
-.039	-.061	.037	.094	.046	-.026	-.013	-.046	.006	.033	.004	.030	.003	-.018	.010	-.031
.396	1.202	.419	.577	.271	.365	.309	.395	.202	.286	.128	.476	.091	.204	.096	.172**
1.598		.996		.636		.704		.488		.604		.295		.268***	
2.594				1.340				1.092				.563****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

Table 5-4

Error Estimates in
$\hat{Y}_{10}$*

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
..009	.108	.029	-.000	-.019	-.007	.029	.003	-.006	.043	-.009	.046	.052	.030	-.009	-.029
-.243	.199	.044	.043	-.052	-.134	-.013	.046	-.058	-.023	-.014	-.004	-.012	.027	-.007	-.045
-.027	-.176	.005	-.033	-.010	.055	.025	.001	-.012	-.004	.014	-.006	.013	-.043	.020	.013
.051	.139	-.011	.062	.026	-.008	-.035	-.022	.013	.030	-.012	-.021	-.024	.006	.009	-.006
-.025	.014	-.012	-.122	-.008	.030	.000	-.053	.012	.046	-.005	-.047	.013	.006	-.002	.008
.056	.011	-.027	.024	-.021	-.051	.024	.011	-.007	.011	.005	-.001	.012	-.008	-.022	.022
.001	-.020	.021	-.011	-.013	-.056	-.006	-.008	.000	.004	-.002	-.001	-.011	-.013	-.017	.018
-.030	.068	.030	.020	-.040	.004	.034	-.091	.002	.007	.014	.066	.002	-.005	.010	-.031
.442	.735	.179	.315	.189	.345	.166	.235	.110	.168	.075	.192	.139	.138	.096	.172**
1.177		.494		.534		.401		.278		.267		.277		.268***	
1.671				.935				.545				.545****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

Table 5-5

Error Estimates in
$\hat{Y}_{11*}$

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
-.082	-.056	.031	.044	-.043	-.058	-.007	.000	-.042	.016	.000	-.012	.006	.040	.004	.015
.054	.138	.038	-.040	.038	.048	-.018	-.022	-.043	-.024	.011	-.013	.060	-.031	.015	-.003
.032	.030	-.007	.005	.105	.010	-.023	.029	-.030	.040	.001	.013	.029	.018	-.002	.014
-.007	.080	-.004	.039	-.009	.051	-.045	-.053	-.001	.027	-.007	.026	.018	.015	-.004	-.021
.011	.103	-.004	.104	-.011	-.001	-.005	.006	.031	-.009	.001	-.017	.000	.026	-.010	.006
-.034	-.044	.013	.015	-.033	-.013	-.001	.012	.003	-.039	.006	.005	.003	-.014	.005	.004
.028	.072	-.006	-.047	-.017	-.011	.025	.031	.004	-.003	.004	.026	.019	-.004	.008	.006
-.003	-.097	-.020	-.058	.048	-.042	-.025	.133	.011	.113	.010	-.040	-.006	-.017	-.010	.005
.251	.620	.123	.352	.304	.234	.149	.286	.165	.271	.040	.152	.141	.165	.058	.074**
.871		.475		.538		.435		.436		.192		.306		.132***	
1.346				.973				.628				.438****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

evaluate the effects of different levels of K and n on these error estimates. Within each level of K, error decreased as n increased. For example, with K = 10, 30, 60 and 150, the absolute error in $\hat{\gamma}_{00}$ dropped from 1.53, .883, .388 and .317 to .989, .646, .355 and .252, respectiviely. This downward trend was consistent over all four $\underset{\sim}{\gamma}$'s. Again the main difference across $\underset{\sim}{\gamma}$ was one of degree. Increasing the number of groups substantially improved the results. Absolute error for all four parameters decreased with larger number of groups. Although the patterns are similar, reduction for $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ are greater than that of $\hat{\gamma}_{10}$ and $\hat{\gamma}_{11}$. Moving from K = 10 to K = 150 reduces the error in $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ to almost 1/5 whereas this reduction is only 1/3 for $\hat{\gamma}_{10}$ and $\hat{\gamma}_{11}$. Among the three factors the effect of K on error of estimates is more pronounced than that of n and d.

These downward trends in d and n were compounded as number of groups was introduced. For example, with regard to sample size, n, the absolute error in $\hat{\gamma}_{00}$ dropped only minimally within each level of K (e.g., going from 1.53, .883, .388, and .317 to .989, .646, .355, and .252, respectively). Increasing the number of groups brought substantial improvement in results. That is, the absolute error dropped from 1.53 under K = 10 to .252 under K = 150. A similar pattern is present in the case of d. While absolute errors of estimate decrease as d decreases under a given level of n (e.g., dropping from 1.013, .558, .276 and .172 to .517, .325, .112 and .145) this reduction is compounded across

levels of K i.e., going from 1.013 under K = 10 to as low as .145 under K =150.

Figures 5-1 and 5-2 provide schematic representations of error patterns in the macro parameters (symbols A-Z and * signify frequencies 10-36, respectively, in the figures used throughout the thesis). In these figures errors are plotted against varying combinations of K and n when d = .10. The plots follow the patterns described earlier in this section. The further we move toward K = 150, the more concentrated the errors are around zero. This suggests that ML estimates of macro parameters are asymptotically consistent. The errors are small for all levels of n when K = 150, and for n = 30, 60 and 150 when K = 60, but not so when K = 10 or 30. This suggests that the latter combinations of K and n produce less precise estimates of the macro parameters. When intraclass correlation is increased to .25 (Figures A-1 and A-2 in Appendix) the exact same pattern is reproduced except the spread of errors is magnified .

Results of multiple regression analysis applied to the squared errors of the macro estimators support the foregoing discussion of the differential effects of K, n and d. Specifically when $(Kn)^{-1}$ and d were the assumed predictors in the regression equations there was a significant linear relationship between the squared errors and the two predictors with $(Kn)^{-1}$ accounting for more variability. That is, when $(Kn)^{-1}$ was the sole predictor, R^2 for γ_{00}, γ_{01}, γ_{10}

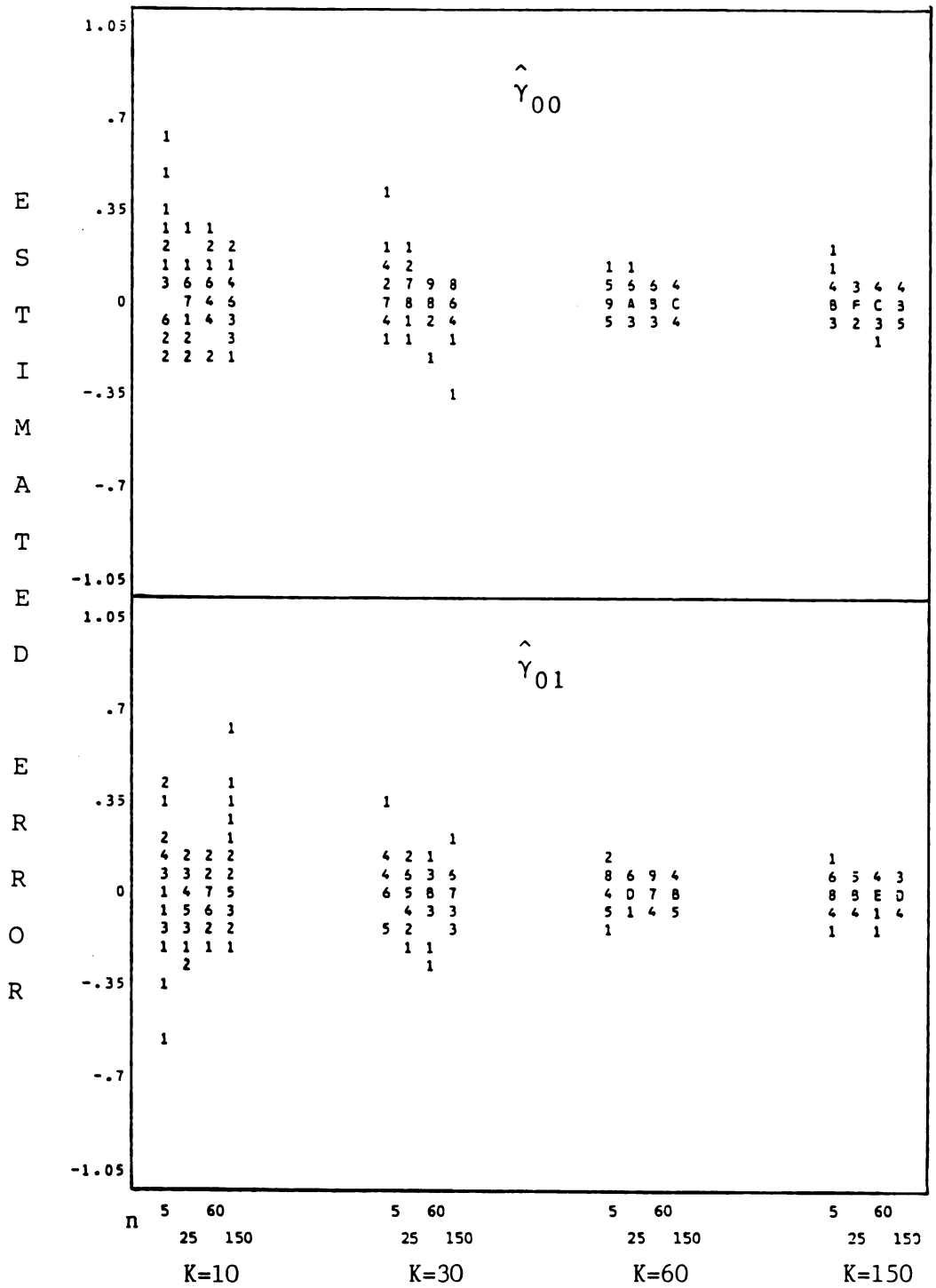

Figure 5-1. Error estimates in $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ for different combinations of k and n with d=.10.

Note: Symbols A-Z signify frequencies 10-35, respectively.

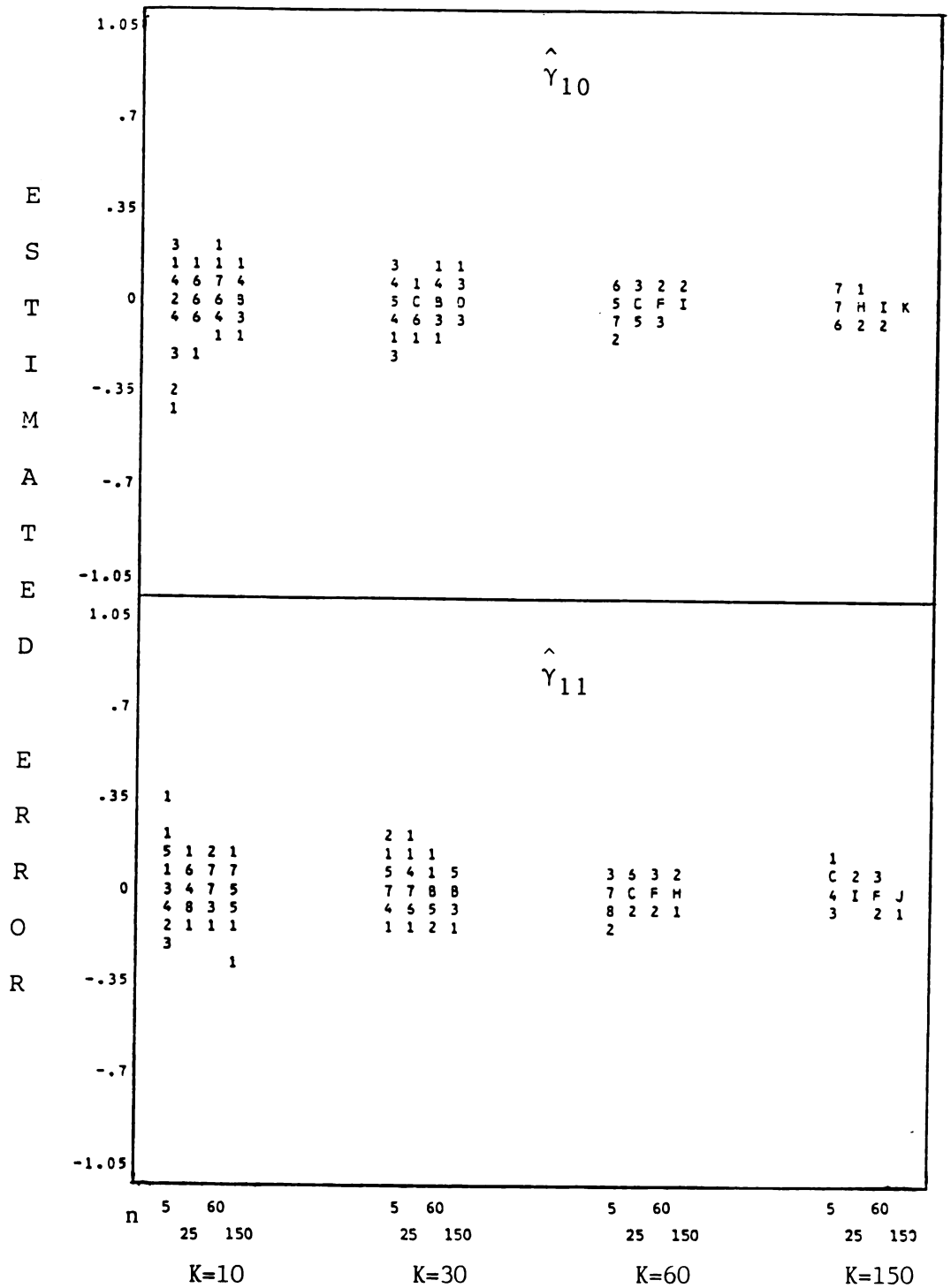

Figure 5-2. Error estimates in $\hat{\gamma}_{10}$ and $\hat{\gamma}_{11}$ for different combinations of k and n with d=.10.

Note: Symbols A-Z signify frequencies 10-35, respectively.

and γ_{11} assumed the values of .11, .15, .19, and .11, respectively and increased to .13, .17, .21, and .13, respectively when d was added to the equation. This finding was consistent across all macro estimators.

However, when the effects of K and n were considered separately, K^{-1} always accounted for more variability followed by d and lastly by n^{-1} (using stepwise regression analysis) except in the case of γ_{10} where K^{-1} was followed first by n^{-1} and then by d.

Are The Macro Parameters Asymptotically Efficient ?

Three estimates of dispersion in macro parameters are computed. These estimates as defined in Chapter IV are, VAR, MSE, and MSW. Table 5-6 contains the mean value of these three statistics averaged over the entire sample along with the asymptotic dispersion of $\hat{\underset{\sim}{\gamma}}^{*}$, the Cramer Rao lower bound (CRLB).

Across all macro parameters the three measures of dispersion are not significantly different from each other and from CRLB. Within each dispersion measure, the magnitude of dispersion for γ_{10} and γ_{11} is smaller than that of γ_{00} and γ_{01}. This pattern is consistent with previously established results. In two occasions MSW gave smaller estimates than CRLB (.007872 and .008636 versus .008058 and .009595 respectively) which can be attributed to sampling variance.

Table 5-6

Measures of Dispersion in Macro Parameters*

	γ_{00}	γ_{01}	γ_{10}	γ_{11}
VAR	.018283	.021039	.009079	.010753
MSW	.017085	.021947	.007872	.008636
MSE	.018341	.021904	.008525	.008377
CRLB	.016640	.019662	.008058	.009594

*From 640 replications. Dispersion measures are: HLM estimates via EM algorithm, VAR; mean squares within, MSW; mean square error, MSE; and Cramer-Rao lower bound, CRLB.

The values of MSW and MSE are very similar. This lends credence to the notion that the estimates are unbiased since the only difference between MSW and MSE is that one uses the mean of the parameter estimates and the other uses the parameter itself. Similarly, the resemblance in values of VAR and CRLB lends itself to the notion that estimates of macro parameters are asymptotically efficient. VAR is different from CRLB in that it uses the ML estimates of variance components and not the parameter variances.

To evaluate the differential effects of K, n and d on VAR and to assess explicitly the discrepancy between VAR and CRLB, Tables 5-7 through 5-9 provide differences between VAR and CRLB (DIF = VAR - CRLB) as a function of K, n, and d , respectively.

The differences portray the extent to which VAR approximates CRLB under varying levels of these three factors . While larger d consistently increased the asymptotic variance and its ML estimates, the difference

Table 5-7

Differences in Measures of Dispersion in Macro Parameters Estimated by HLM via EM Algorithm (VAR) and Cramer-Rao Lower Bound (CRLB) for Different Number of Groups*

	K	VAR	CRLB	VAR-CRLB
γ_{00}	10	.238891	.221223	.017668
	30	.075480	.066246	.009234
	60	.036600	.032476	.004124
	150	.014689	.012854	.001835
γ_{01}	10	.288704	.276495	.012209
	30	.078131	.068933	.009198
	60	.039283	.034971	.004312
	150	.014664	.012844	.001820
γ_{10}	10	.124528	.108462	.016066
	30	.034335	.031260	.003075
	60	.016307	.015386	.000921
	150	.006406	.006052	.000354
γ_{11}	10	.155503	.136727	.018776
	30	.036303	.033094	.003209
	60	.016776	.015934	.000842
	150	.006479	.006126	.000353

*From 160 replications of K groups.

Table 5-8

Differences in Measures of Dispersion in Macro Parameters Estimated by HLM via EM Algorithm (VAR) and Cramer-Rao Lower Bound (CRLB) for Different Group Sizes*

	n	VAR	CRLB	VAR-CRLB
	5	.138723	.108442	.030281
γ_{00}	25	.091868	.088070	.008061
	60	.065642	.065254	.000388
	150	.069427	.075295	-.005868
	5	.149417	.126111	.023306
γ_{01}	25	.099376	.092726	.006650
	60	.076937	.076393	.000544
	150	.095051	.098014	-.002961
	5	.090942	.073461	.017481
γ_{10}	25	.040522	.037811	.002711
	60	.020843	.021008	-.000165
	150	.029270	.028880	.000390
	5	.104826	.085723	.019103
γ_{11}	25	.046579	.042289	.004290
	60	.027006	.026608	.000398
	150	.036650	.037262	-.000612

*From 160 replications of groups of size n.

Table 5-9

Differences in Measures of Dispersion in Macro Parameters Estimated by HLM via EM Algorithm (VAR) and Cramer-Rao Lower Bound (CRLB) for Different Intraclass Correlation Coefficients*

	d	VAR	CRLB	VAR-CRLB
γ_{00}	.10	.053569	.044934	.008635
	.25	.129261	.121465	.007796
γ_{01}	.10	.064928	.055877	.009031
	.25	.145462	.140745	.004717
γ_{10}	.10	.031157	.025678	.005479
	.25	.059631	.054902	.004729
γ_{11}	.10	.036040	.031199	.004841
	.25	.071491	.064741	.004081

*From 320 replications of d intraclass correlation coefficient.

between the two estimates, although not statistically significant, is reduced. This downward trend is considerably more pronounced with an increase in K and consistent among all macro parameters. Results for n, while generally following the same trend, did not appear to be smallest with the largest sample size. Under all four macro parameters, the smallest difference occurred with n = 60. It might be assumed that this result was due to the pattern of factor combinations in matrix design, since n = 60 only occurs in "block 0" and n = 150 in "block 1". Additionally, the values of VAR and CRLB consistently get smaller proportionally with an increase in K and decrease in d . This pattern is not followed when n increased. This might be due to the fact that parameter variances $\tau_{\mu|W}$ and $\tau_{\beta|W}$ are affected by K and not by n. However, the situation is reversed with σ^2 which is affected by sample size and not by so much number of groups. But considering that σ^2 is fixed at 1 and is not allowed to vary explains the pronounced effect of K over n in reducing the size of the VAR and CRLB.

Schematically these close approximations between VAR and CRLB are presented for γ_{00}, γ_{01} , γ_{10} and γ_{11} in Figures 5-3 and 5-4 respectively. This is done by first transforming VAR and CRLB as such:

$\sqrt{K/2}$*log (VAR) + 1/K and $\sqrt{K/2}$*log (CRLB), respectively. The logarithmic transformation is used in order to stabilize VAR and CRLB. When unstabilized the sampling variance of the statistic, VAR, dependes on the size of CRLB, the parameter

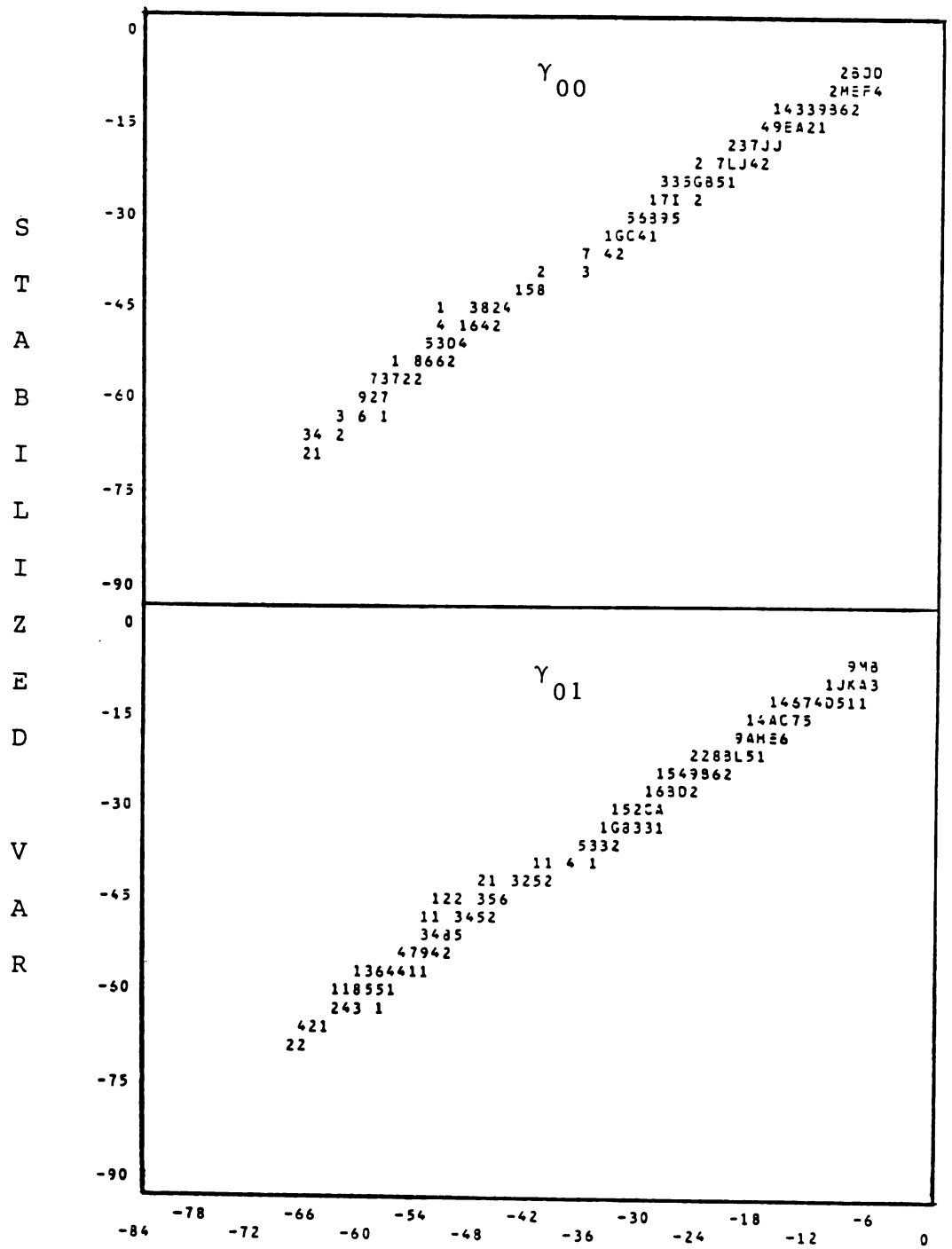

Figure 5-3. Plot of stabilized VAR and CRLB for γ_{00} and γ_{01}.

Note: Symbols A-Z signify frequencies 10-35, respectively.

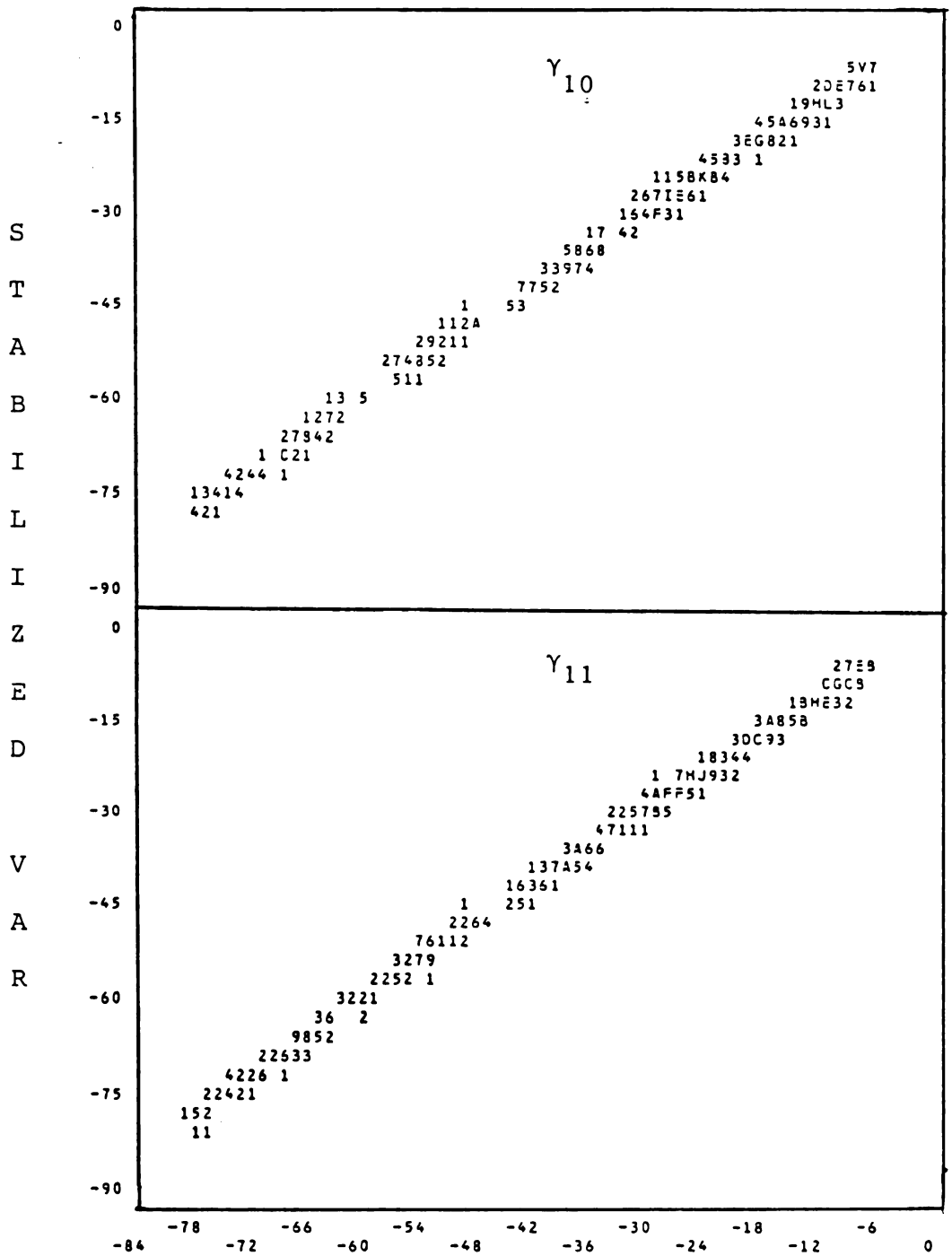

Figure 5-4. Plot of stabilized VAR and CRLB for γ_{10} and γ_{11}.

Note: Symbols A-Z signify frequencies 10-35, respectively.

value. The relationship although not perfect, is highly predictable. The slope of regression lines is over .985 for γ_{00} and γ_{01} and over .999 for γ_{10} and γ_{11}. The intercepts are -.03 and .13 respectively.

The preceding discussion of differential effects of K and n and d is supported by applying multiple regression analysis to the DIF squared calculated for each macro estimator. Specifically, when $(Kn)^{-1}$ and d were the assumed predictors, there was a significant linear relationship between DIF squared and the two predictors with $(Kn)^{-1}$ consistently accounting for more variability. This finding was consistent across all macro estimators.

However, when the effect of K and n were considered separately K^{-1} always accounted for more variability followed by d^{-1}. n did not account for more variability in (DIF) above and beyond K^{-1} and d. The only exception was for γ_{00} where R^2 increased from .125 to .134 as a result of introducing n^{-1} to the equation.

It is worth noting that squared multiple correlation (R^2) for the latter set of prediction equations are consistently higher than that of former set of equations. This may be taken as an evidence to the pronounced effect of K over n in reducing the size of dispersion in macro parameters. This is because the second set of equations reflect the effect of K separate from n as is opposed to the combined effects of K and n in the first set of equations.

Do the Macro Parameters Have Asymptotic Normal Distribution?

One of the desirable properties of ML estimates is that their asymptotic distribution is known to be normal. This property indeed facilitates the process of inference making. Macro estimators are distributed as:

$$(\hat{\gamma}^* - \gamma) \overset{ASY}{\sim} N\left| 0, \left(\Sigma W_j' \Delta_j^{-1} W_j \right)^{-1} \right| ,$$

or

$$Z = ((\gamma_h^* - \gamma_h)/S.E. (\gamma_h^*)) \overset{ASY}{\sim} N\left| 0, 1 \right| ,$$

where subscript h refers to the elements in the vector.

This property is examined through normal probability plotting technique applied to the Z statistics computed for each macro parameter. This method applies a special transformation to the vertical scale of a graph of the assumed type cumulative distribution function. As a result, the cumulative distribution of the assumed type will transform into a straight line. If the Z statistics are indeed normally distributed the corresponding quantiles of the model when plotted against the sample order statistics will be nearly linear. When all sample points are considered the sampling distribution of the Z-statistics closely approximate normal distribution over all macro parameters (Figure 5-5). Slight deviations from linearity are due to random sampling fluctuations. Notice that the Z statistics assume different range of values. These ranges are -3 to +4, -4 to +3, -4 to +3, and -4 to +4 for $Z_{\gamma_{00}}$, $Z_{\gamma_{01}}$, $Z_{\gamma_{10}}$, and $Z_{\gamma_{11}}$ rspectively. Table 5-10 contains the mean and variance of the Z statistics.

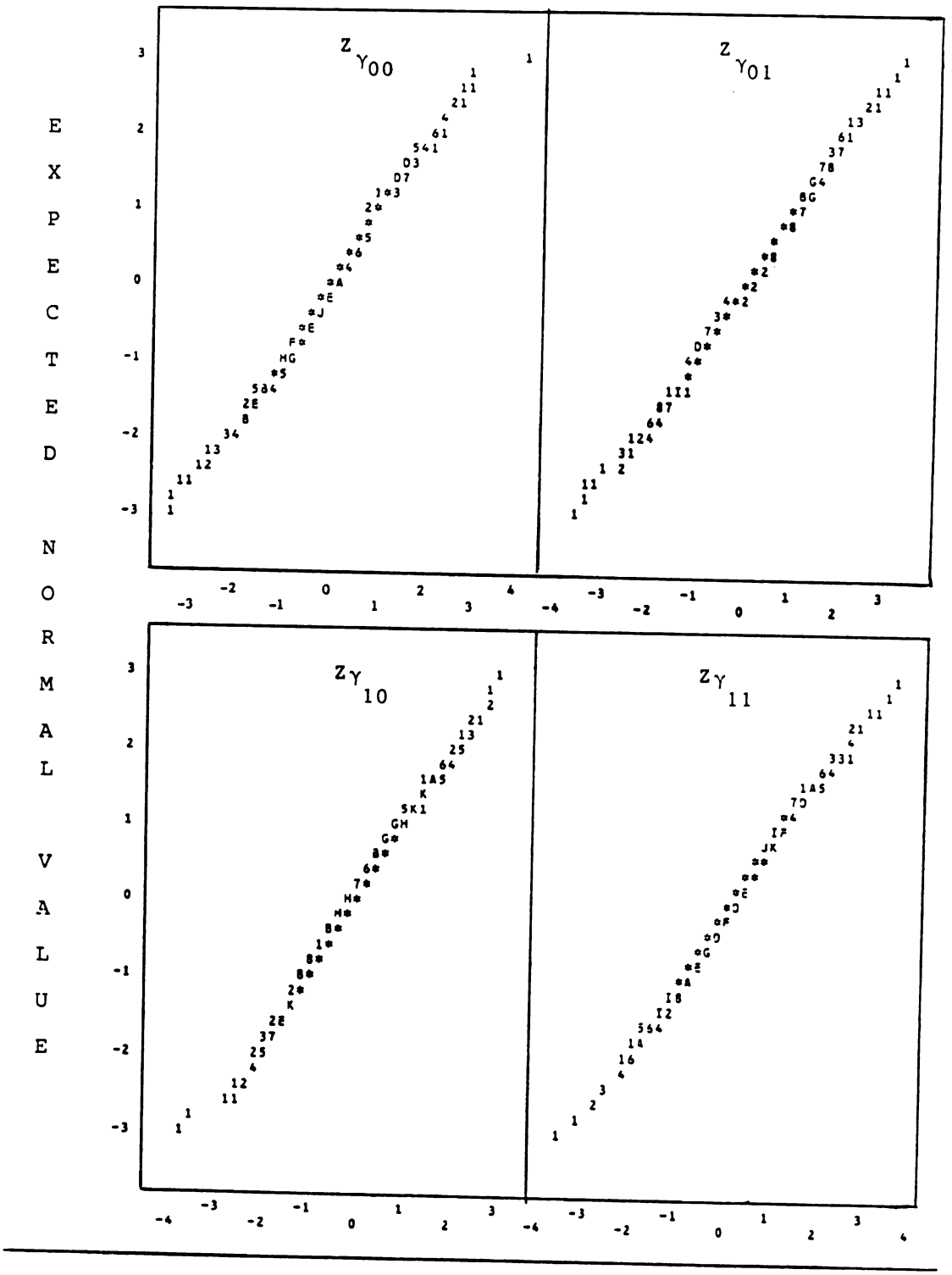

Figure 5-5. Normal probability plot of Z statistics for γ_{00}, γ_{01}, γ_{10}, and γ_{11}.

Note: Symbols A-Z and * signify frequencies 10-36, respectively.

Table 5-10
Mean and Variance of the
Z-Statistics for the Macro Parameters*

	$Z_{\gamma_{00}}$	$Z_{\gamma_{01}}$	$Z_{\gamma_{10}}$	$Z_{\gamma_{11}}$
Mean	.038	.000	-.041	.072
Variance	1.072	1.137	1.047	1.127

*From 640 replications.

The mean and variance of the Z statistics for different number of groups, group sizes, and intraclass correlation coefficients are presented in Tables A-1 through A-3, respectively. No pattern developed between the means or the variances of the Z statistics and the number of groups, group sizes or intraclass correlation coefficients.

Are the Variance Components Asymptotically Unbiased and Consistent ?

The expected errors of the variance components, $\hat{\tau}_{\mu|W}$, $\hat{\tau}_{\beta|W}$ and $\hat{\sigma}^2$ are reported in Tables 5-11 through 5-13, respectively. The cell values are calculated by first deviating the estimated values from their respective parameter values ($\tau_{\mu|W}$, $\tau_{\beta|W}$ and $\sigma^2 = 1$) and then averaging them over the number of replications. The 95 percent confidence interval constructed for the expected errors of variance-covariance components are given below:

95% C.I. for expected error of $\tau_{\mu|W}$ = (.0373, .0647)

95% C.I. for expected error of $\tau_{\beta|W}$ = (.0031, .0149)

95% C.I. for expected error of $\tau_{\mu\beta|W}$ = (-.0089, .0029)

95% C.I. for expected error of σ^2 = (-.0049, .0067)

Table 5-11

Error Estimates in $\hat{\tau}_{\mu|W}$*

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
-.020	.568	-.027	-.278	.070	.193	-.012	-.107	.020	.149	-.014	-.040	.025	.201	.015	.057
-.021	-.053	.092	.072	.306	.029	-.004	-.037	.292	.093	-.023	.039	.286	.046	.026	.005
.059	.011	-.015	.135	.046	.465	-.007	.021	.032	.333	.007	-.009	.017	.253	-.000	-.001
.531	.168	-.006	.093	.506	-.095	.016	.022	.282	.088	.008	-.006	.327	.056	.036	-.007
.084	.016	-.108	-.329	.049	.103	-.019	.035	.023	.234	-.022	.003	.093	.152	.008	.061
.048	-.031	.030	-.009	-.025	-.007	.027	-.060	.005	-.012	.010	.006	.005	.002	-.006	-.011
.085	-.034	-.034	.013	.036	.002	.037	-.020	.052	-.010	.015	-.011	.055	.014	.007	-.016
.002	.115	-.018	.066	-.003	.098	-.003	.140	-.002	.104	-.001	.017	.002	.134	.001	-.000
.850	.996	.330	.995	1.041	.992	.125	.442	.708	1.023	.100	.091	.810	.858	.099	.158**
1.846		1.325		2.033		.567		1.731		.191		1.668		.257***	
3.171				2.600				1.922				1.925****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

Table 5-12

Error Estimates in $\hat{\tau}_{\beta|W}$ *

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
.027	.050	-.009	-.023	.020	.007	.002	.012	.009	.021	.001	.014	-.003	-.011	.004	.007
.158	.095	-.032	-.004	.026	-.083	.020	.025	.029	-.014	.017	-.004	-.023	.029	-.004	.015
.061	.079	.000	-.002	.029	.051	-.000	.003	.043	.015	.001	.001	.010	.015	-.001	-.004
.070	-.033	-.001	.045	.071	.052	-.004	-.011	.025	.008	-.008	-.003	.019	.010	.004	.016
.018	.004	-.003	.132	-.007	-.003	-.000	-.033	.003	-.002	-.000	-.008	-.003	.030	.001	-.005
.009	.005	-.016	-.003	.010	.010	-.002	-.001	.001	.001	.000	.001	-.002	.000	.000	-.000
.037	.017	-.008	-.001	-.005	-.010	.002	.000	.003	.003	.001	-.002	.001	-.000	-.003	-.001
.024	.025	.008	-.121	-.012	.075	.009	.073	.004	-.057	-.002	.016	-.000	-.036	-.000	.036
.404	.308	.077	.331	.180	.291	.039	.158	.117	.121	.030	.049	.061	.131	.017	.084**
.712		.408		.471		.197		.238		.079		.192		.101***	
1.120				.668				.317				.293****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

Table 5-13

Error Estimates in $\hat{\sigma}^2$ *

K=10				K=30				K=60				K=150			
n'		n"		n'		n"		n'		n"		n'		n"	
d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25	d=.10	d=.25
.150	.059	.004	.022	-.012	-.011	.014	-.008	.021	.038	.004	.000	-.028	-.009	.000	.002
-.144	.023	.047	-.014	-.019	.074	.009	.001	-.046	.045	.011	.003	.037	-.038	-.001	-.007
-.037	.023	.016	-.013	-.017	-.069	.013	-.029	-.022	-.046	.010	.001	.023	-.023	-.005	.006
.126	-.041	-.010	.003	-.047	.119	.004	-.008	.007	-.078	.015	.001	.025	.003	-.007	.006
.015	.027	-.005	.005	.018	-.004	-.014	-.018	.003	-.001	-.005	.017	-.016	-.014	-.013	.000
-.005	.024	-.011	.012	.021	.008	-.001	-.003	-.034	-.014	.002	.002	-.007	-.020	-.003	.001
-.037	.019	-.001	.005	.007	-.018	.010	.015	.014	.011	.000	-.001	.005	-.015	-.002	-.003
-.037	-.022	-.011	.002	-.004	.001	.001	-.003	.016	.021	.005	-.004	.001	-.012	-.002	-.003
.551	.238	.105	.076	.145	.304	.066	.085	.163	.254	.052	.029	.142	.134	.033	.028**
.784		.181		.449		.151		.417		.081		.276		.061***	
.970				.600				.498				.337****			

*From 5 replications of K groups of size n'(n=5 or 25) and n"(n=60 or 150) with d intraclass correlation coefficient
**Absolute errors of estimate summed within levels of d
***Absolute errors of estimate summed within levels of n
****Absolute errors of estimate summed within levels of k

The expected errors of $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$ are significantly different from zero i.e., $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$ are biased estimators of $\tau_{\mu|W}$ and $\tau_{\beta|W}$ respectively. However, their covariance estimator $\hat{\tau}_{\mu\beta|W}$ is an unbiased estimator of $\tau_{\mu\beta|W}$ (note: $\tau_{\mu\beta|W}$ is assumed to be zero in model specification section). Finally, $\hat{\sigma}^2$ may considered to be an unbiased estimator of σ^2 since zero is included in the 95% confidence interval of its expected errors. Basically the error estimates follow a similar general pattern to that observed for the macro parameters. That is, error consistently decreased as either n or K or both increased. However the intraclass correlation coefficient did not produce a consistent effect across all 16 cells.

Also the size of the errors in $\hat{\tau}_{\beta|W}$ and $\hat{\sigma}^2$ are consistently smaller than that of $\hat{\tau}_{\mu|W}$. This is due to the assumed parameter variance in slope to that of intercept used to generate data which is larger for μ than is for β .

Figures 5-6 and 5-7 demonstrate these patterns graphically. In these plots the variance components are first transformed logarithmically, corrected for bias and then plotted against varying combinations of K and n with d = .10 . This transformation has reversed the range of the error of estimates in $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$. As a result $\hat{\tau}_{\beta|W}$ now has larger range of errors than $\hat{\tau}_{\mu|W}$. Even though $\hat{\sigma}^2$ relies more on n than K, it behaves better when the compound effects of K and n are present. For example, when d = .10

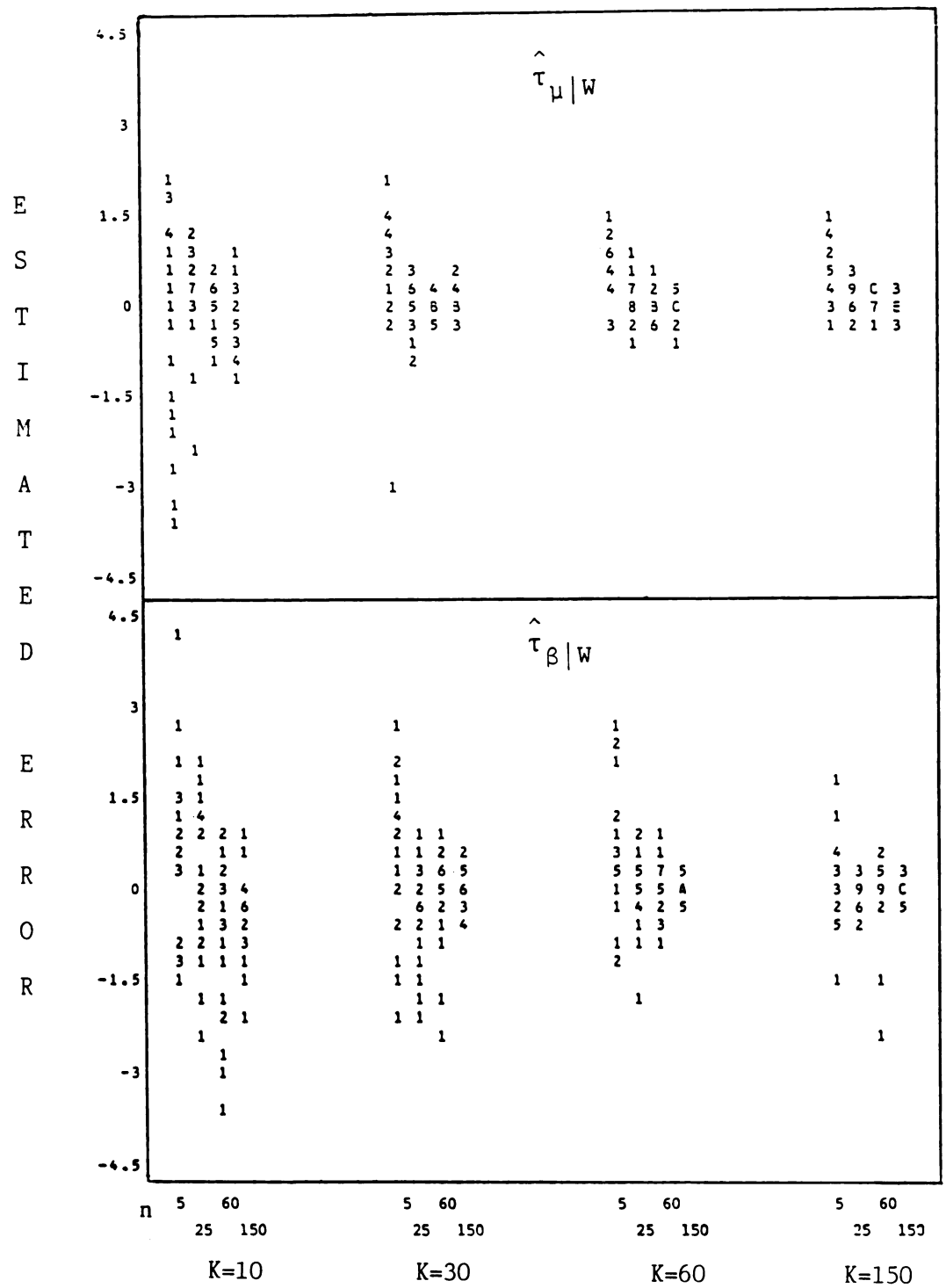

Figure 5-6. Error estimates in transformed $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$ for different combinations of k and n with d=.10.

Note: Symbols A-Z signify frequencies 10-35, respectively.

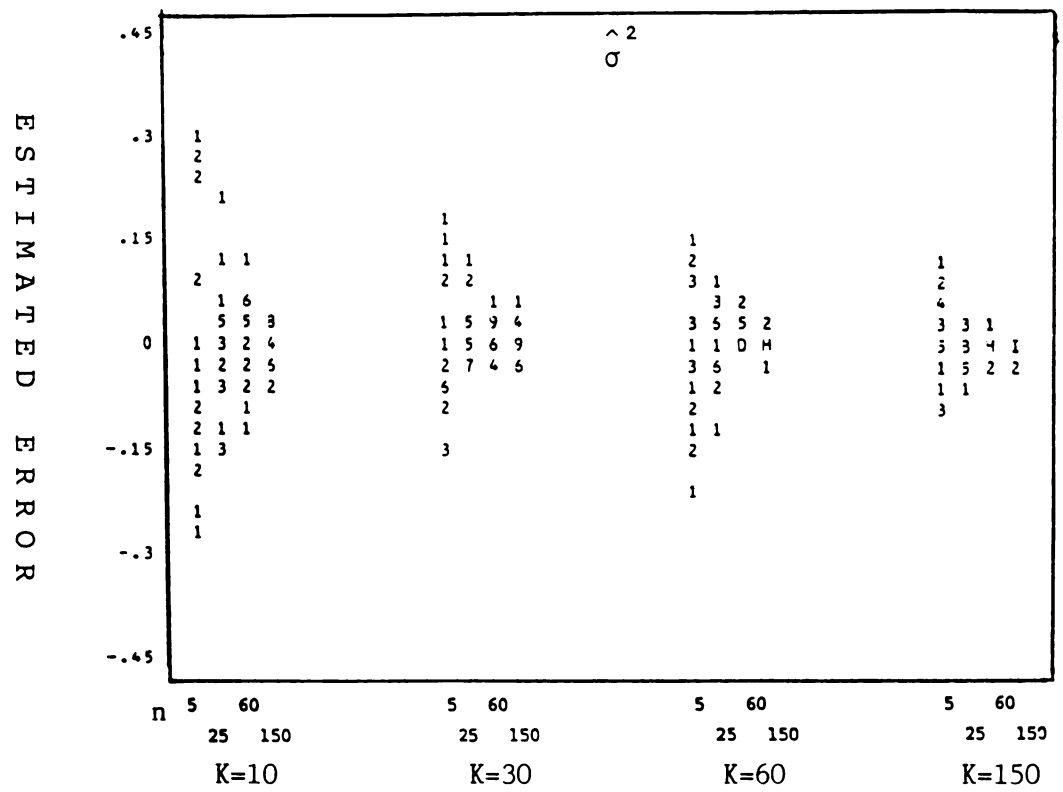

Figure 5-7. Error estimates in transformed $\hat{\sigma}^2$ for different combinations of k and n with d=.10.

Note: Symbols A-Z signify frequencies 10-35, respectively.

with n = 150, $\hat{\sigma}^2$ is reliable even when K = 10. But when K increases to 30, or more, $\hat{\sigma}^2$ becomes reliable even with n = 60.

When the intraclass correlation is increased to .25 the similar pattern is reproduced except the spread of errors is generally reduced (Figures A-3 and A-4 in Appendix). Notice that this adverse effect of intraclass correlation coefficient is in contrast to its effect on macro parameters discussed earlier.

Once again when multiple regression analysis is applied to the squared errors of these transformed variance components the foregoing discussion of differential effects of K, n, and d is supported. Specifically, a significant linear relationship is evidenced between the squared errors of $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$ and $(Kn)^{-1}$. The intraclass correlation coefficient d, with negative slope across all variance components, is significant only for $\hat{\tau}_{\mu|W}$.

When the effect of K and n are considered separately n^{-1} accounted for more variability in squared error of $\hat{\tau}_{\mu|W}$ followed by K^{-1} and then d. However, reverse happened for $\hat{\tau}_{\beta|W}$. That is, K^{-1} explained the most variability followed by n^{-1}, but with no significant contribution from d above and beyond the two predictors.

When σ^2 was considered, the predictor n^{-1} acounted for the most variability followed by K^{-1}. Furthermore, when the effect of these predictors were combined into $(Kn)^{-1}$, R^2 increased from .25 to .31. However, intraclass correlation

coefficient did not contribute to either case.

Are the Variance Components Asymptotically Efficient ?

This property is examined graphically through plotting the squared of the transformed values of the variance components against their respective asymptotic variance (Figures 5-8 and 5-9). The asymptotic variance of $\hat{\tau}_{\mu|W}$, $\hat{\tau}_{\beta|W}$ and σ^2 are 2/K, 2/K and 2/nK, respectively. $\hat{\sigma}^2$ approximates its asymptotic variance quite well, as n and K increase. The estimates of $\hat{\sigma}^2$ seem to be reliable for all combinations of K and n except when both take the lowest value i.e., n = 5 and K = 10.

The two parameter variances approximate their respective asymptotic variances quite well except when K = 10. However, across all levels of K estimator $\hat{\tau}_{\mu|W}$ gives a better approximation of its asymptotic variance than that of $\hat{\tau}_{\beta|W}$.

Results for the Hypothesis Testing Phase

The objective for this portion of the study is to assess the effect of variance estimation on inferences about macro parameters with respect to both robustness and power. To the extent that variance components are misestimated, the estimated macro coefficients and their standard errors will be in error thus affecting the inferences made about macro parameters.

The situation considered is that of four different

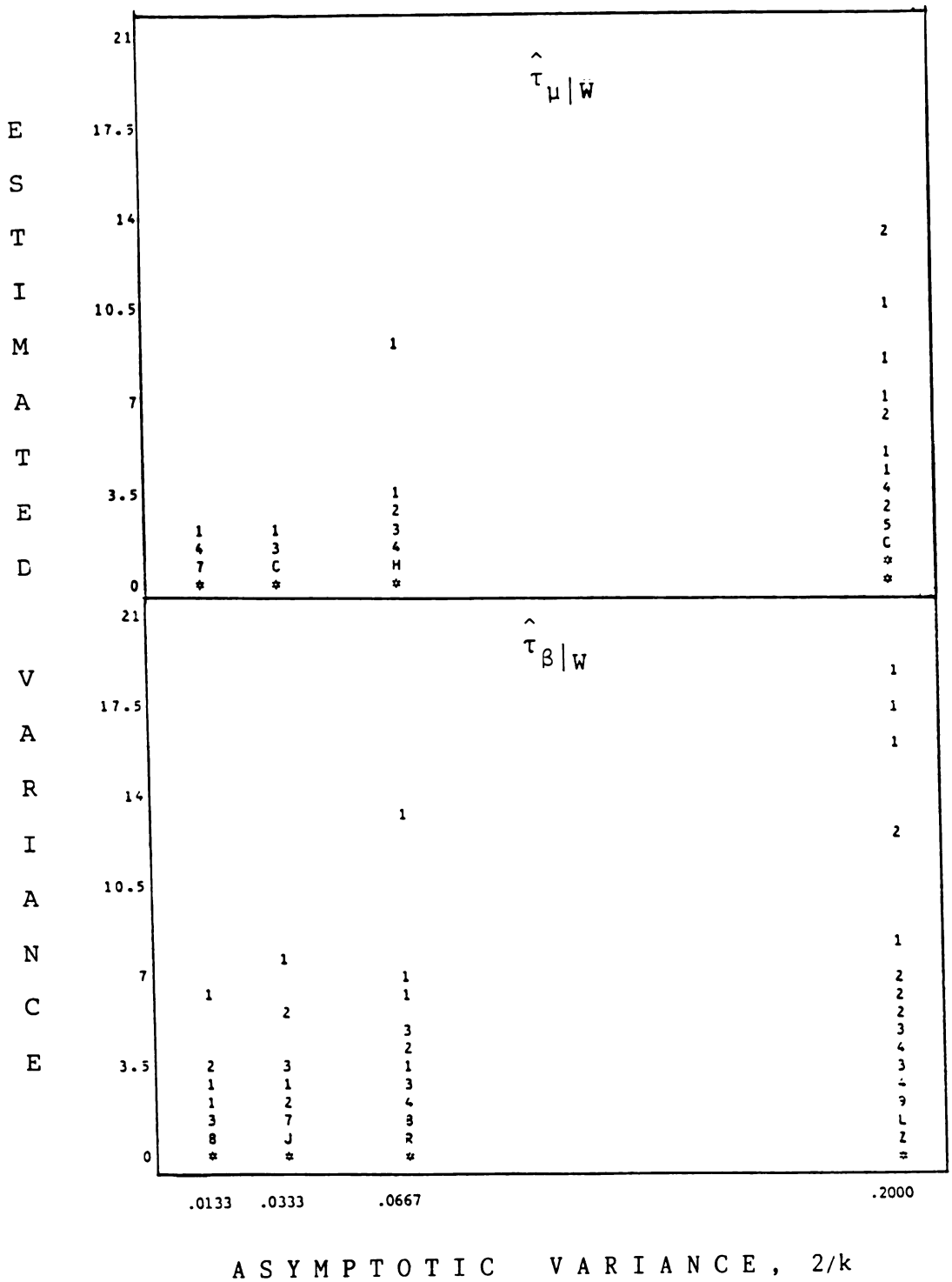

Figure 5-8. Plot of transformed estimated and asymptotic variance of $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$.

Note: Symbols A-Z and * signify frequencies 10-36, respectively.

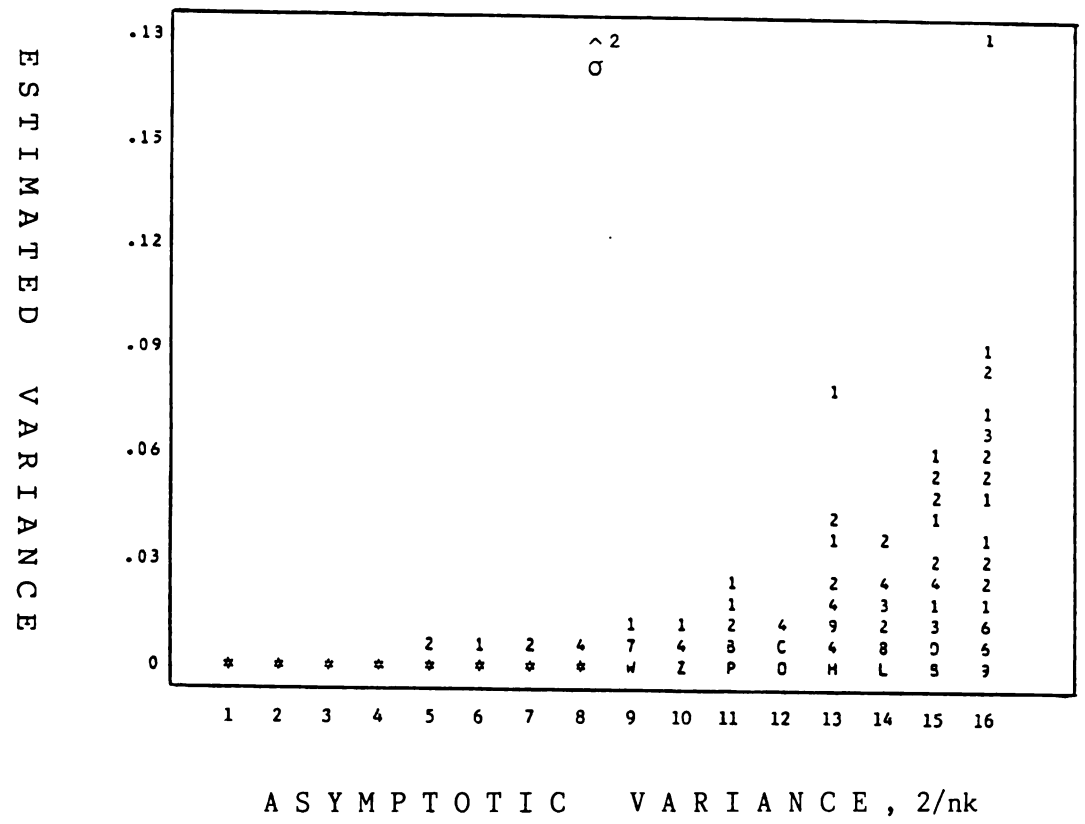

Figure 5-9. Plot of transformed estimated and asymptotic variance of $\hat{\sigma}^2$, Where:

$1=2/n_4k_4$	$2=2/n_3k_4$	$3=2/n_4k_3$	$4=2/n_4k_2$
$5=2/n_2k_4$	$6=2/n_3k_3$	$7=2/n_3k_2$	$8=2/n_4k_1$
$9=2/n_2k_3$	$10=2/n_2k_2$	$11=2/n_1k_4$	$12=2/n_3k_1$
$13=2/n_1k_3$	$14=2/n_2k_1$	$15=2/n_1k_2$	$16=2/n_1k_1$

$k_1=10$	$k_2=30$	$k_3=60$	$k_4=150$
$n_1=5$	$n_2=25$	$n_3=60$	$n_4=150$

Note: Symbols A-Z and * signify frequencies 10-36, respectively.

combinations of K and n along with five other factors described in Chapter IV. The design for this part of the study allows for an assessment of robustness and power under unknown variance components when: (1) total sample size is considered, N; (2) the number of groups is varied (K = 10, 30, 60, and 150); (3) group size is varied (n = 5, 25, 60, and 150); and (4) the intraclass correlation coefficient is varied (d = .10, and .25).

The data for the first part consists of 640 replications, parts two and three contain 160 replications each of four combinations of K and n, and the last part is based on 320 replications each of two combinations of d.

Robustness Under Various Conditions

This section evaluates the effect of variance estimation on tests of macro parameters based on total sample size, N, different levels of K, n, and intraclass correlation coefficient, d. Since the data are randomly generated via Monte Carlo methods, random error in the data must be considered. To take this error into account, the standard error (S.E.) of a proportion for a sample size equal to the number of replications is employed. The S.E. for a proportion is estimated by $(P(1 - P)/N)^{1/2}$, where P is the true value of the proportion, and N equals the number of replications. Since the true value of P (i.e., nominal alpha) is known, this formula is used to calculate the S.E. at the three nominal alpha levels considered. These are given in Table 5-14.

Table 5-14

Standard Errors for Nominal Alpha Levels
and Number of Replications Used in the Study

Alpha	N=640	N=320	N=160
.01	.0039	.0056	.0079
.05	.0086	.0122	.0172
.10	.0119	.0168	.0237

With a reduction in the number of generated data sets comes an increase in standard errors. Given known parameters (i.e., nominal alpha levels), the standard error of a proportion may be used to calculate confidence intervals around the known parameters instead of probability intervals around the sample estimates. Using the standard procedure, 95 and 99 percent confidence intervals for the three nominal levels considered are presented in Table 5-15. Thus, obtained alpha levels within these intervals may be considered to be within sampling error of nominal alpha.

Total Sample Size and Robustness

Table 5-16 contains the actual alpha levels for all parameters under central unknown variance components situation when all combinations of K and n considered together. The empirical type I error rates are consistently exceeding the nominal error rates across all macro

Table 5-15

Probability Intervals for Nominal Alpha Levels and Number of Replications Used in the Study

a) 95% Probability Intervals

Alpha	N=640	N=320	N=160
.01	(.0024, .0176)	(.0000, .0210)	(.0000, .0255)
.05	(.0331, .0669)	(.0261, .0739)	(.0163, .0837)
.10	(.0767, .1233)	(.0671, .1329)	(.0535, .1465)

b) 99% Probability Intervals

Alpha	N=640	N=320	N=160
.01	(.0000, .0201)	(.0000, .0245)	(.0000, .0304)
.05	(.0278, .0722)	(.0185, .0815)	(.0056, .0944)
.10	(.0693, .1307)	(.0567, .1433)	(.0389, .1612)

Table 5-16

Type I Error Rates for Tests of Macro Estimators Under a True Null*

	e.s.	α =.01	α =.05	α =.10
γ_{00}	0	.020**	.059	.119
γ_{01}	.31	.022**	.066	.119
γ_{10}	.76	.013	.061	.112
γ_{11}	.16	.019**	.072**	.122

*From 640 replications with e.s. effect size.

**Outside the 95% confidence interval.

parameters. Especially the error rates are relatively large when testing γ_{00}, γ_{01} and γ_{11} than that of γ_{10}. However, most values tend to be within 95% confidence intervals. The exceptions are for γ_{00} and γ_{01} at α = .01, and for γ_{11} at .01 and .05 alpha where all are within 99% confidence interval but γ_{01}. When outside the probability intervals, empirical alpha levels are all liberal.

Number of Groups and Robustness

Tables 5-17 through 5-20 (part a) present the type I error rates for tests of macro estimators γ_{00}, γ_{01}, γ_{10} and γ_{11}, respectively when the number of groups varied, with K = 10, 30, 60, and 150. The values for all macro parameters tend to be within 95% confidence intervals of the nominal alpha across all K levels. When outside the confidence interval, empirical significance levels are all liberal. Exceptions are for γ_{10} with K = 30 at .01 and .05 alpha and for γ_{11} with K = 30 and 150 at .05, .10 and .05 alpha respectively. However, these values are typically within 99% confidence interval. An unexpected finding from this set of results is that for a given macro parameter the largest type I error occurred randomly regardless of number of groups.

Sample Size and Robustness

Tables 5-17 through 5-20 (part b) give the type I error rates for tests of macro parameters γ_{00}, γ_{01}, γ_{10}, and γ_{11} respectively for experimental conditions with n = 5, 25, 60,

Table 5-17

Type I Error Rates for Test of Macro Parameter, γ_{00}, Under a True Null

a) For different number of groups, k.*

k	e.s.	α =.01	α =.05	α =.10
10	0	.025	.075	.119
30		.019	.075	.138
60		.013	.031	.087
150		.025	.056	.131

*From 160 replications with e.s. effect size.

b) For different group size, n.*

n	e.s.	α =.01	α =.05	α =.10
5	0	.025	.063	.112
25		.019	.050	.119
60		.019	.050	.119
150		.019	.075	.125

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

d	e.s.	α =.01	α =.05	α =.10
.10	0	.022**	.047	.100
.25		.019	.072	.138

*From 320 replications with e.s. effect size.

**Outside the 95% confidence interval.

Table 5-18

Type I Error Rates for Tests of Macro Parameter, γ_{01}, Under a True Null

a) For different number of groups, k.*

k	e.s.	α =.01	α =.05	α =.10
10	.31	.025	.081	.138
30		.019	.056	.112
60		.025	.050	.100
150		.019	.075	.125

*From 160 replications with e.s. effect size.

b) For different group size, n.*

n	e.s.	α =.01	α =.05	α =.10
5	.31	.031**	.075	.138
25	.32	.013	.050	.081
60	.31	.031**	.081	.156**
150	.32	.013	.056	.100

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

d	e.s.	α =.01	α =.05	α =.10
.10	.22	.019	.053	.109
.25	.41	.025**	.078**	.128

*From 320 replications with e.s. effect size.

**Outside the 95% confidence interval.

Table 5-19

Type I Error Rates for Tests of Macro Parameter, γ_{10}, Under a True Null

a) For different number of groups, k.*

k	e.s.	α =.01	α =.05	α =.10
10	.76	.013	.044	.106
30		.031**	.094**	.119
60		.006	.056	.106
150		.000	.050	.119

*From 160 replications with e.s. effect size.

b) For different group size, n.*

h	e.s	α =.01	α =.05	α =.10
5	.73	.006	.050	.100
25	.78	.006	.075	.144
60	.73	.025	.081	.125
150	.78	.013	.038	.081

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

d	e.s	α =.01	α =.05	α =.10
.10	.72	.019	.059	.103
.25	.79	.006	.063	.122

*From 320 replications with e.s. effect size.

**Outside the 95% confidence interval.

Table 5-20

Type I Error Rates for Tests of Macro Parameter, γ_{11}, Under a True Null

a) For different number of groups, k.*

k	e.s.	α =.01	α =.05	α =.10
10	.16	.019	.069	.112
30		.025	.087**	.150**
60		.019	.038	.087
150		.013	.094**	.138

*From 160 replications with e.s. effect size.

b) For different group size, n.*

n	e.s.	α =.01	α =.05	α =.10
5	.16	.006	.044	.081
25	.17	.031**	.069	.106
60	.16	.013	.081	.138
150	.17	.025	.094**	.162**

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

d	e.s.	α=.01	α=.05	α=.10
.10	.11	.025**	.081**	.128
.25	.22	.013	.063	.116

*From 320 replications with e.s. effect size.

**Outside the 95% confidence interval.

and 150. For all macro parameters, actual significance levels tend to be within the 95% probability intervals of nominal values across all levels of n. When outside the confidence intervals, empirical alpha levels are all liberal. Exceptions are for γ_{01} with n = 5 at .01 alpha and with n = 60 at .01 and .10 alpha, and for γ_{11} with n = 25 at .01 alpha and with n = 150 at .05 and .10 alpha levels. However, all values are typically within 99% probability intervals of the nominal alpha. Again no pattern emerged. That is, for a given macro parameter the largest type I error occurred randomly regardless of sample size and effect size. The only exception is for γ_{01} where departure from nominal alpha was fairly small with an increase in effect size with n = 25 and 150.

Intraclass Correlation Coefficient and Robustness

Tables 5-17 through 5-20 (part c) report the type I error rates for tests of macro estimators γ_{00}, γ_{01}, γ_{10}, and γ_{11}, respectively when the intraclass correlation coefficient varied, with d = .10 and .25. Once again most values tend to be within 95% confidence intervals of the nominal alpha for both levels of d. The values outside of this interval are all liberal. However, values not contained within 95% confidence intervals are all within 99% probability interval. These values are: γ_{00} with d = .10 at α = .01; γ_{01} with d = .25 at α = .01 and .05; and γ_{11} with d = .10 at α = .01 and .05. Again no pattern emerged

with respect to d and /or effect sizes.

Power Under Various Conditions

The goal of this portion of the study is to evaluate the power of the tests of macro parameters in rejecting the null hypothesis under unknown variance components situation by considering total sample size, N, different levels of K, n and intraclass correlation coefficient, d. The empirical estimates of power (P') may also be compared to the theoretical values of power (P") obtained through nominal noncentrality parameter discussed in Chapter IV.

Because of the way the null and true alternative hypotheses are set up (see Chapter IV) the implementation and discussion of power analysis will be limited only to macro parameters γ_{01} , γ_{10} , and γ_{11} . The actual parameter value of γ_{00} is set at zero. Thus, power analysis cannot be applied.

Total Sample Size and Power

As shown in Table 5-21 the empirical power for all macro parameters considering all four combinations of K and n together are quite high.

Of the three macro parameters γ_{10} consistently obtains the highest power, followed by γ_{01} and γ_{11} . This pattern of order in power is highly consistent with the magnitude of the effect sizes in the macro parameters. Within each macro parameter power is always larger at larger nominal levels.

Table 5-21

Power for Tests of Macro Parameters*

		α =.01		α =.05		α=.10	
	e.s.	p'	p"	p'	p"	p'**	p"***
γ_{01}	.31	.9382	.9495	.9846	.9881	.9932	.9949
γ_{10}	.76	.9999	.9999	.9999	.9999	.9999	.9999
γ_{11}	.16	.8023	.7995	.9292	.9265	.9625	.9616

*From 640 replications with e.s. effect size.
**Empirical power
***Nominal power

The power estimates of macro parameters are either equal or very close to the theoretical values (differences are statistically insignificant). Within each nominal alpha level, empirical power is smaller than the nominal power for γ_{01} . The situation is reversed with respect to γ_{11} .

Number of Groups and Power

Tables 5-22 through 5-24 (part a) give empirical and nominal power for macro parameters γ_{01} , γ_{10} , and γ_{11} , respectively when the number of groups varied, with K = 10, 30, 60, and 150. Again γ_{10} has the highest power across all levels of K and all alpha levels. γ_{01} with effect size e.s. = .31 obtains the next highest place but reaches the same degree of power (.9999) with K = 150 at all levels of alpha. On the other hand, γ_{11} obtains the same power with the same value of K but only at α = .05 and .10.

Table 5-22

Power for Tests of Macro Parameter

γ_{01}

a) For different number of groups, k.*

		α =.01		α =.05		α =.10	
k	e.s.	p'	p"	p'	p"	p'**	p"***
10	.31	.2546	.2119	.4801	.4286	.6064	.5517
30		.6736	.7054	.8577	.8770	.9162	.9292
60		.9686	.9772	.9934	.9955	.9974	.9983
150		.9999	.9999	.9999	.9999	.9999	.9999

*From 160 replications with e.s. effect size.

b) For different groups size, n.*

		α =.01		α =.05		α =.10	
n	e.s.	p'	p"	p'	p"	p'	p"
5	.31	.6141	.7517	.8186	.9015	.8888	.9463
25	.32	.9406	.9564	.9854	.9898	.9936	.9959
60	.31	.9772	.9778	.9956	.9957	.9984	.9984
150	.32	.9893	.9846	.9982	.9973	.9993	.9989

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

		α =.01		α =.05		α =.10	
d	e.s.	p'	p"	p'	p"	p'	p"
.10	.22	.8997	.9265	.9713	.9808	.9864	.9913
.25	.41	.9641	.9664	.9920	.9927	.9968	.9971

*From 320 replications with e.s. effect size.

**Empirical power
***Nominal power

Table 5-23

Power for Tests of Macro Parameter
γ_{10}

a) For different number of groups, k.*

		α=.01		α=.05		α=.10	
k	e.s.	p'	p"	p'	p"	p'**	p"***
10	.76	.9999	.9999	.9999	.9999	.9999	.9999
30		.9999	.9999	.9999	.9999	.9999	.9999
60		.9999	.9999	.9999	.9999	.9999	.9999
150		.9999	.9999	.9999	.9999	.9999	.9999

*From 160 replications with e.s. effect size.

b) For different group size, n.*

		α =.01		α =.05		α =.10	
k	e.s.	p'	p"	p'	p"	p'	p"
5	.73	.9999	.9999	.9999	.9999	.9999	.9999
25	.78	.9999	.9999	.9999	.9999	.9999	.9999
60	.73	.9999	.9999	.9999	.9999	.9999	.9999
150	.78	.9999	.9999	.9999	.9999	.9999	.9999

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

		α =.01		α =.05		α =.10	
d	e.s.	p'	p"	p'	p"	p'	p"
.10	.72	.9999	.9999	.9999	.9999	.9999	.9999
.25	.79	.9999	.9999	.9999	.9999	.9999	.9999

*From 320 replications with e.s. effect size.

**Empirical power
***Nominal power

Table 5-24

Power for Tests of Macro Parameter

γ_{11}

a) For different number of groups, k.*

		α =.01		α =.05		α =.10	
k	e.s.	p'	p"	p'	p"	p'**	p"***
10	.16	.1423	.1314	.3264	.3050	.4443	.4247
30		.5040	.4840	.7357	.7190	.8264	.8159
60		.8665	.8621	.9582	.9564	.9798	.9783
150		.9996	.9997	.9999	.9999	.9999	.9999

*From 160 replications with e.s. effect size.

b) For different group size, n.*

		α =.01		α =.05		α =.10	
n	e.s.	p'	p"	p'	p"	p'	p"
5	.16	.2743	.3156	.5040	.5517	.6293	.6736
25	.17	.7422	.7357	.8962	.8944	.9429	.9406
60	.16	.9162	.9082	.9767	.9744	.9896	.9881
150	.17	.9767	.9693	.9955	.9936	.9982	.9974

*From 160 replications with e.s. effect size.

c) For different intraclass correlation coefficients, d.*

		α =.01		α =.05		α =.10	
d	e.s.	p'	p"	p'	p"	p'	p"
.10	.11	.6915	.6808	.8686	.8621	.9236	.9207
.25	.22	.8849	.8849	.9656	.9656	.9834	.9834

*From 320 replications with e.s. effect size.

** Empirical power
***Nominal power

As shown in Figure 5-10 with K = 150 the power curves for the three macro parameters are indistinguishable. With respect to number of groups, power is best with K = 60 and 150 and worst with K = 10 when γ_{01} is considered and best with K = 150 and worst with K = 10 for γ_{11} .

Empirical powers are always close but smaller than the theoretical powers for γ_{01} except with K = 10 and close but larger than the nominal power for γ_{11} except with K = 150 (differences between P' and P" are atatistically insignificant).

Sample Size and Power

Tables 5-22 through 5-24 (part b) give tha actual and nominal power for macro parameters γ_{01}, γ_{10} , and γ_{11} , respectively for groups of size 5, 25, 60 and 150. Across all levels of n power improved relative to what it was when levels of K was considered. This was consistent across all macro parameters and three alpha levels (to evaluate simultaneous effect of K and n on empirical and theoretical power see Tables A-4 through A-6 in Appendix). γ_{10} the macro parameter with the larger effect size gained the most power. With each macro parameter power increased as either effect size, group size, or alpha level increased.

As shown in Figure 5-11 again with n = 150 the power curves for the three macro parameters are indistinguishable. With regard to sample size, power is best with n = 25 or above for γ_{01} , but only with n = 60 and 150 when γ_{11} is considered.

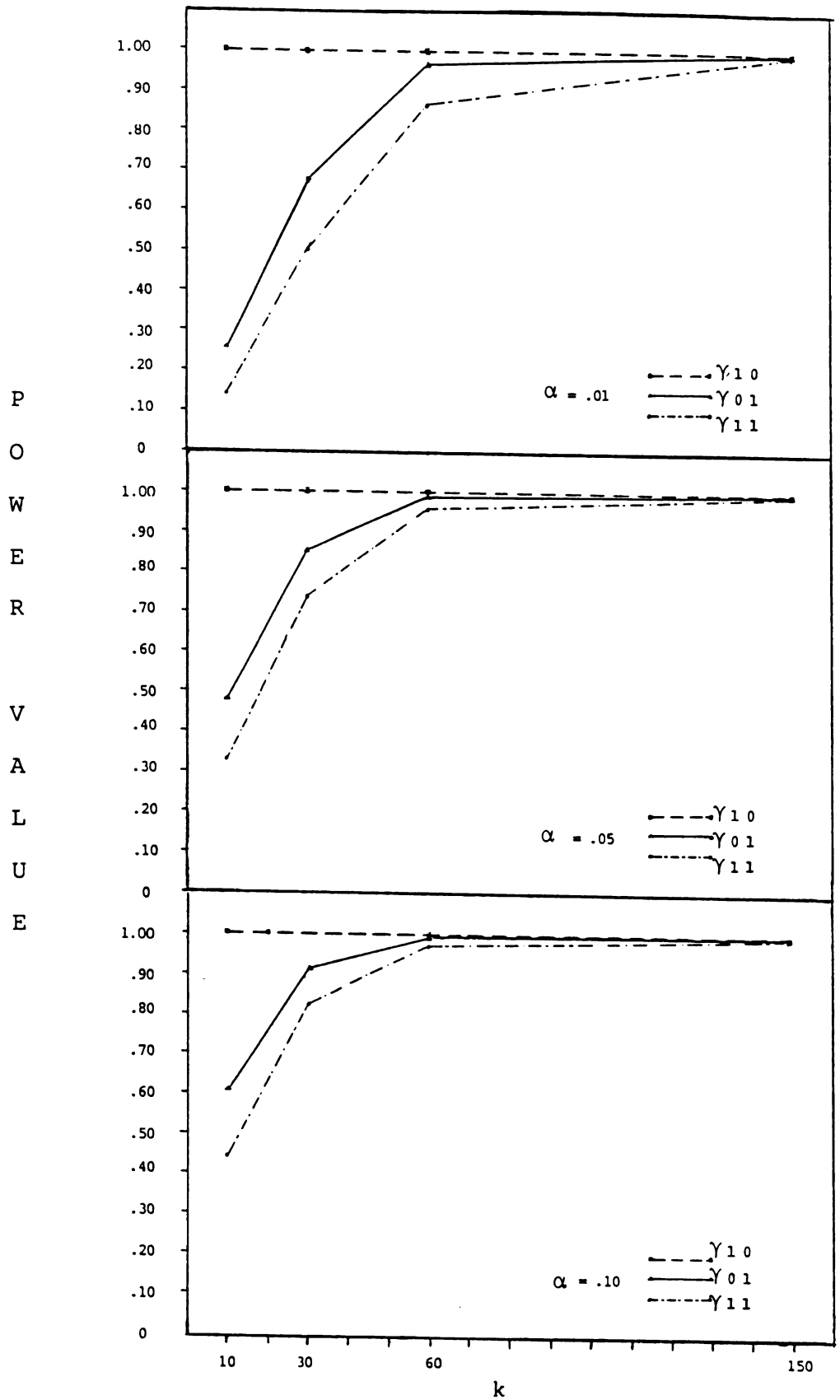

Figure 5-10. Power curves of γ_{01}, γ_{10} and γ_{11} for different no. of groups, k.

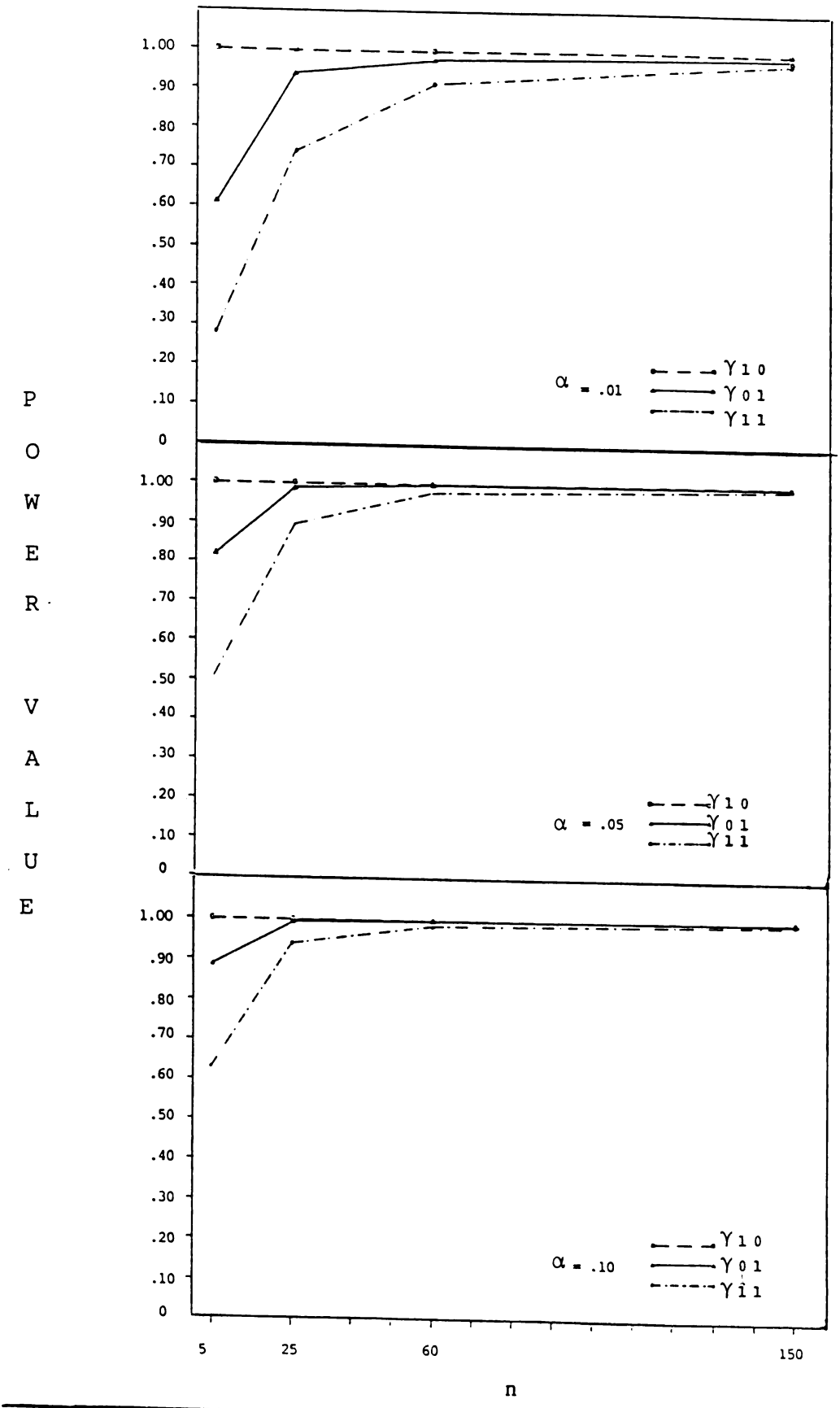

Figure 5-11. Power curves of $\gamma 01$, $\gamma 10$ and $\gamma 11$ for different group sizes, n.

Although statistically insignificant nominal power is larger than empirical power for γ_{01} except when n = 150. The situation is reversed when γ_{11} is considered. That is, empirical power always exceeds theoretical power except with n = 5 .

Intraclass Correlation Coefficient and Power

Part c of Tables 5-22 through 5-24 reflects the effect of intraclass correlation coefficient on power of the macro parameters γ_{01}, γ_{10} and γ_{11} , respectively. Within each macro parameter power increased as intraclass correlation coefficient increased. This upward trend was consistent across all macro parameters. But this is an artifact due to the effect size. Notice that across all macro parameters larger values of effect size are paired with the larger values of intraclass correlation coefficients. Obviously, the highest power is attained by γ_{10} with the largest effect size. Deviations from nominal power was consistently small and statistically insignificant for all macro parameters. Tables A-7 through A-12 in Appendix represent differential effect of K, n, and d on empirical and nominal powers of the macro parameters.

EM Algorithm: Rate of Convergence

Often EM algorithm is praised for its simplicity of implementation and numerical stability. Nonetheless, its rate of convergence is criticized to be slow. A concern here with this algorithm is its rate of convergence under varying

combinations of K and n used in this study.

The convergence criterion was set at .0001 with the maximum number of iterations allowed fixed at 500. This limit was reached only three times out of 640; two occasions with K = 10 and n = 5, and one occasion with K = 30 and n = 5.

Table 5-25 reports average convergence rate of EM algorithm with respect to sixteen different combinations of K and n.

Table 5-25

Average Convergence Rate of
EM Algorithm

	K=10	K=30	K=60	K=150
n=5	259.125	148.750	116.750	106.875
n=25	49.750	19.250	15.625	7.875
n=60	17.875	6.625	3.875	2.125
n=150	3.625	2.125	2.000	2.000

There is a clear downward trend in average number of iterations prior to convergence to ML estimates both with regard to K and n . Within each level of n the average

number of iterations consistently reduced with an increase in K. This reduction is even more dramatic as a result of n increasing within any level of K. This is an evidence to the pronounced effect of n over K in reducing the rate of convergence in EM algorithm. However, any combination of K = 30, 60 and 150 with n = 60 and 150 results in a reasonably small number of iterations and even more so when n = 150 is combined with any level of K.

CHAPTER VI

DISCUSSION

The results presented in the previous chapter provide an indication that maximum likelihood estimators of macro parameters generated by the EM algorithm behaved as expected. Conclusions based on these results will be presented in this chapter, followed by guidelines for the researcher analyzing multi level data and suggestions for future research.

Conclusions

Under the conditions considered in this study, it appears that:

(1) Macro coefficients generated by the EM algorithm are asymptotically unbiased and consistent. However, as a result of prefixing parameter $c = \tau_\beta/\tau_\mu$ to values .10 and .50 errors in $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ estimates tended to be slightly higher than that of $\hat{\gamma}_{10}$ and $\hat{\gamma}_{11}$. This reflected itself in dispersion measure where again was higher among $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$. Additionally, error estimates are considerably affected by the number of groups and that their precision is directly proportional to K. But, group size (n) does not have a consistent effect on the precision, whereas intraclass correlation coefficient is proportionally related to the squared error of the macro parameters.

(2) Macro coefficients also tend to be asymptotically efficient. Their variances when stabilized closely approximate Cramer Rao minimum variance bound. Slopes of regression lines are over .98 across all macro coefficients.

(3) As far as variance components are concerned $\hat{\sigma}^2$ is extremely well behaved, while the two parameter variances tend to have slightly positive bias. Within group variance $\hat{\sigma}^2$ is considerably affected by group size. However, an increase in the number of groups will farther reduce the estimated errors in $\hat{\sigma}^2$. On the other hand, $\hat{\tau}_{\mu|W}$ is only influenced by group size, and $\hat{\tau}_{\beta|W}$ by number of groups. In contrast to macro parameters precision in the variance components is directly related to the intraclass correlation coefficient. All three variance components approximated their asymptotic variance quite well.

(4) Additionally, test statistics for all macro coefficients resemble normal curve except with slightly heavier tails. This manifests itself in a larger empirical error rates which fall beyond two standard errors from the nominal alpha levels in less than half of the times.

(5) Furthermore, these tests maintain reasonable power under unknown variance components, except for small sample sizes. Specifically, macro coefficient γ_{10} with the largest effect size has the lead followed by γ_{01} and γ_{11} , respectively. Notice that γ_{01} despite having a lower precision obtains a higher power over γ_{11} across all levels of K and n. This is the case since γ_{01} has an effect size which is

double the size of that for γ_{11} .

Guidelines for the Researcher

The analysis of multilevel data may be undertaken with a focus on estimating either micro parameters or macro parameters or both. From the perspective of shrinkage estimation, macro coefficients are merely vehicles for improving estimation of micro effects. However, macro models themselves may be of great importance primarily because they enrich the class of research questions asked about educational effects.

Design of research to optimize allocation of resources is contingent upon the goal and scope of the study. It is helpful to consider three different cases. In the first case, suppose the interest is focused on the estimation of micro parameters at school level with little or no concern for a broader scale. With a large enough data gathered from within a particular school (micro unit) micro parameters can be estimated with reasonable precision i.e., sampling variance V will be near zero. On the other hand, the micro parameters can be estimated by utilizing data not only from that particular school but from many similar schools . In this situation macro coefficients become a secondary concern merely used for improving estimation of micro parameters through empirical Bayes shrinking estimator. On the other extreme, suppose the scope of the study goes beyond one school system and has a potential of becoming a national policy. This immediately entails consideration of

more than one group at the price of reducing the within groups allocation. In this case, estimation of macro models are of primary and only concern. Notice that situations two and three reflect different conceptualizations of macro models. Nonetheless, as far as estimation of macro coefficients is concerned the two situations are identical.

Since the focus of the present study is on the macro parameters, only situations two and three will be addressed. Either situation is concerned with the same question: How can research resources be allocated for the sake of estimating macro coefficients ? Specifically, how many subjects should be sampled from within groups (n) and how many groups (K) should be selected ?

Obviously, determination of statistical power is of primary concern as a preinvestigation procedure. Findings from this study provide strong evidence that the tests of macro parameters maintain reasonable power under unknown variance components, except for small sample sizes. Tables A-4 through A-6 in Appendix may be consulted for an expected power value for different combinations of K and n . Furthermore, Tables A-7 through A-9 in Appendix provide power values for different combinations of K and n along with effect size for macro parameters in the population. However, allocation of samples within and between macro units would not be optimal unless implemented through maximization of power for a given cost. That is, selection of any combination of K and n for an expected power should be done by

minimizing total cost.

Assume the design that has been adopted for a particular study calls for the subjects to be selected from within micro units (e.g. schools) with moderately low intraclass correlation coefficients (i.e., d = .10). Additionally assume that the effect size is approximately equal to .10. The experimenter now needs to consider the same question of optimal allocation of samples within and between units. First consideration should be given to the cost of data collection per group and per subject within a group. The second consideration pertains to the power of the test. The researcher may want a test which is powerful enough to detect even small differences if they exist (e.g. .80 or better). By utilizing information contained in Tables A-6 through A-12 the following combinations of K and n would result in power of at least .81 for tests of macro parameters:

1)	2)	3)
K = 30	K = 60	K = 150
n = 150	n = 25	n = 5

Obviously, selection of any of these combinations is determined by the relative cost of a unit of K and n, and by a given budget allocation. On the other hand, with an improvement in intraclass correlation coefficient of .25 and of an effect size of approximately .20 one can obtain the same magnitude of power for tests of macro parameters even with K = 30 and n = 60. It is worth noting that the suggested n in each set indicates the minimum allocation.

Assuming more budget, this can be exceeded to obtain even a higher power.

As far as the robustness is concerned the results from this study, although of preliminary nature, indicate that macro coefficients' tests are all liberal. Thus, inferences about macro parameters should be made with great caution.

Suggestions for Future Research

The guidelines just discussed are based on a standardized two-stage hierarchical linear model with unit normal predictors at both levels of hierarchy. An aspect that needs to be considered although it is quite likely that this would not produce any radical changes in the results is to let both predictors assume general normal distributions. Moreover, consideration should be given to multiple predictors at each stage of hierarchy, thereby allowing interaction effect to be studied as well. Also, further research should let the within group variance matrix Σ assume a general matrix form. In the present study it was assumed that Σ is a diagonal matrix with equal diagonal elements.

Additionally, the intraclass correlation coefficient on micro level predictor X would be a relevant factor to be considered. Homogeneity of groups on X assumed by the present study reduces the mixed model to being a fixed effect model. But in cluster sampling physical distribution of the population is generally not random but characterized by some homogeneity. Thus, fixing this coefficient at zero

produces an unrealistic situation.

Although findings from this study are of a preliminary nature, they provide strong evidence that, even for small number of replications, tests of macro parameters are reasonably powerful under unknown variance components, except for small sample sizes. Nonetheless to substantiate or refute the results of the present study with regard to robustness to estimated variance components, consideration should be given to a large scale study of a similar nature. Given a larger number of replications, it is expected to achieve some improvements in robustness.

A P P E N D I X

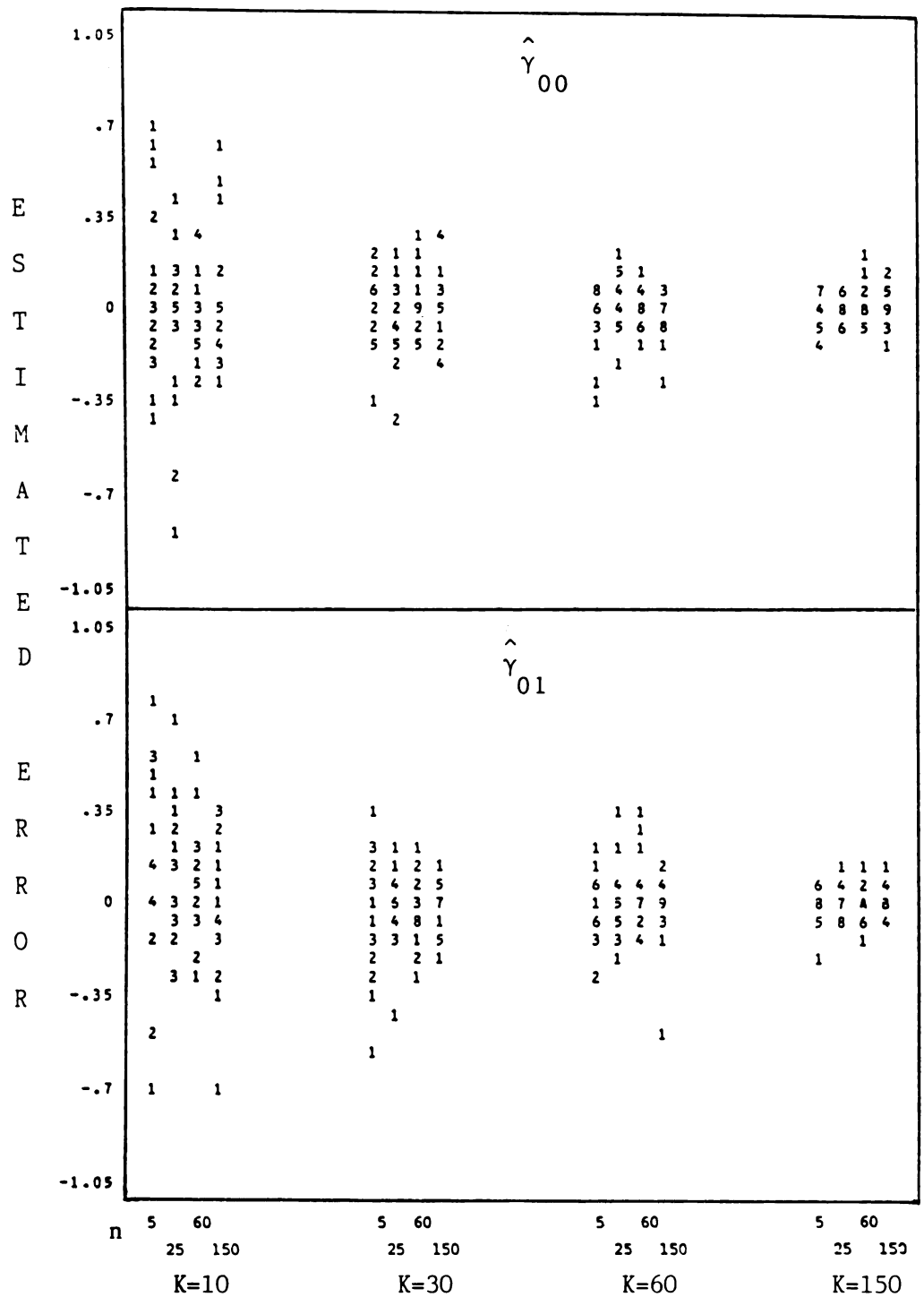

Figure A-1. Error estimates in $\hat{\gamma}_{00}$ and $\hat{\gamma}_{01}$ for different combinations of k and n with d=.25.

Note: Symbols A-Z signify frequencies 10-35, respectively.

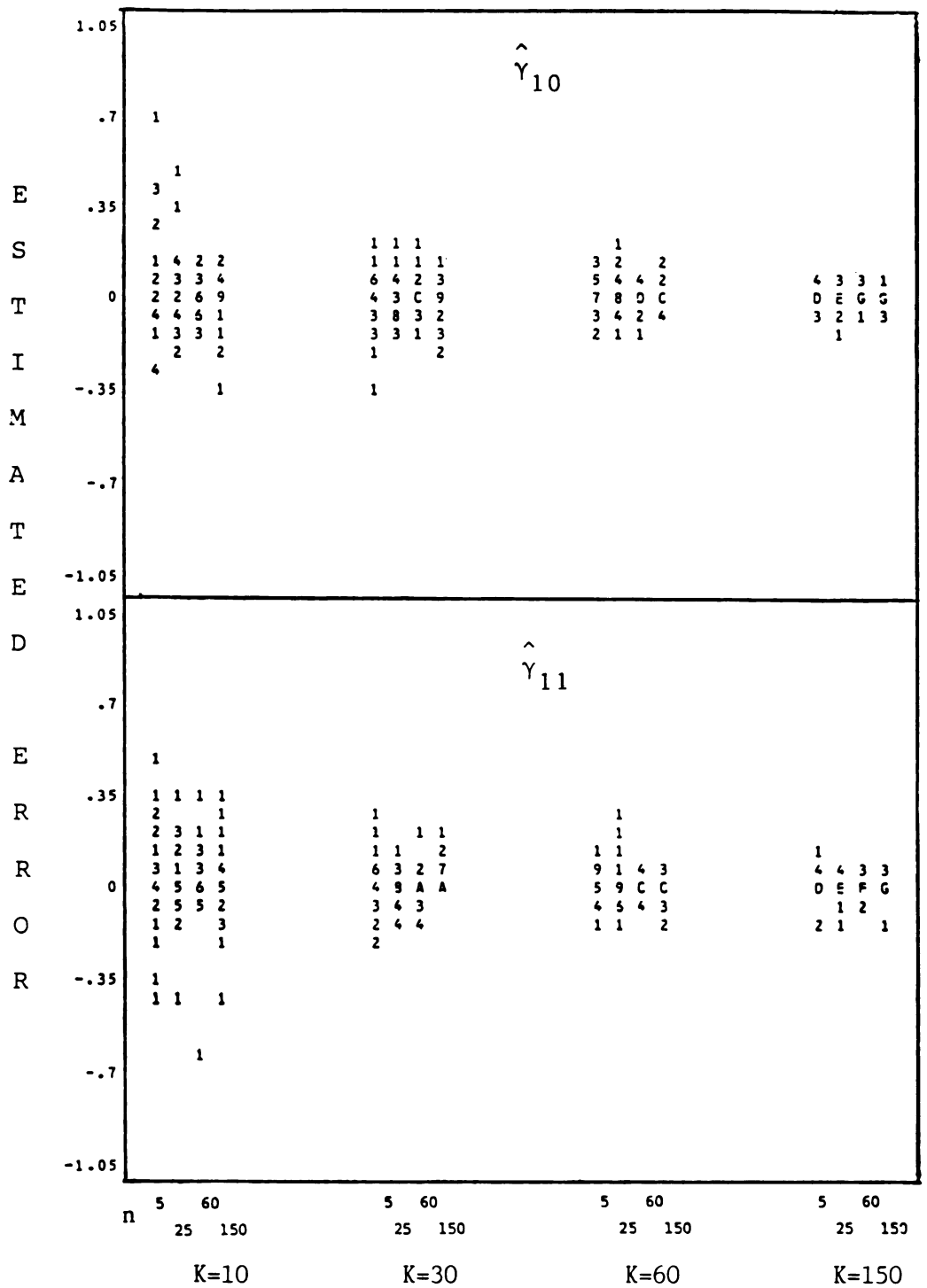

Figure A-2. Error estimates in γ_{10} and γ_{11} for different combinations of k and n with d=.25.

Note: Symbols A-Z signify frequencies 10-35, respectively.

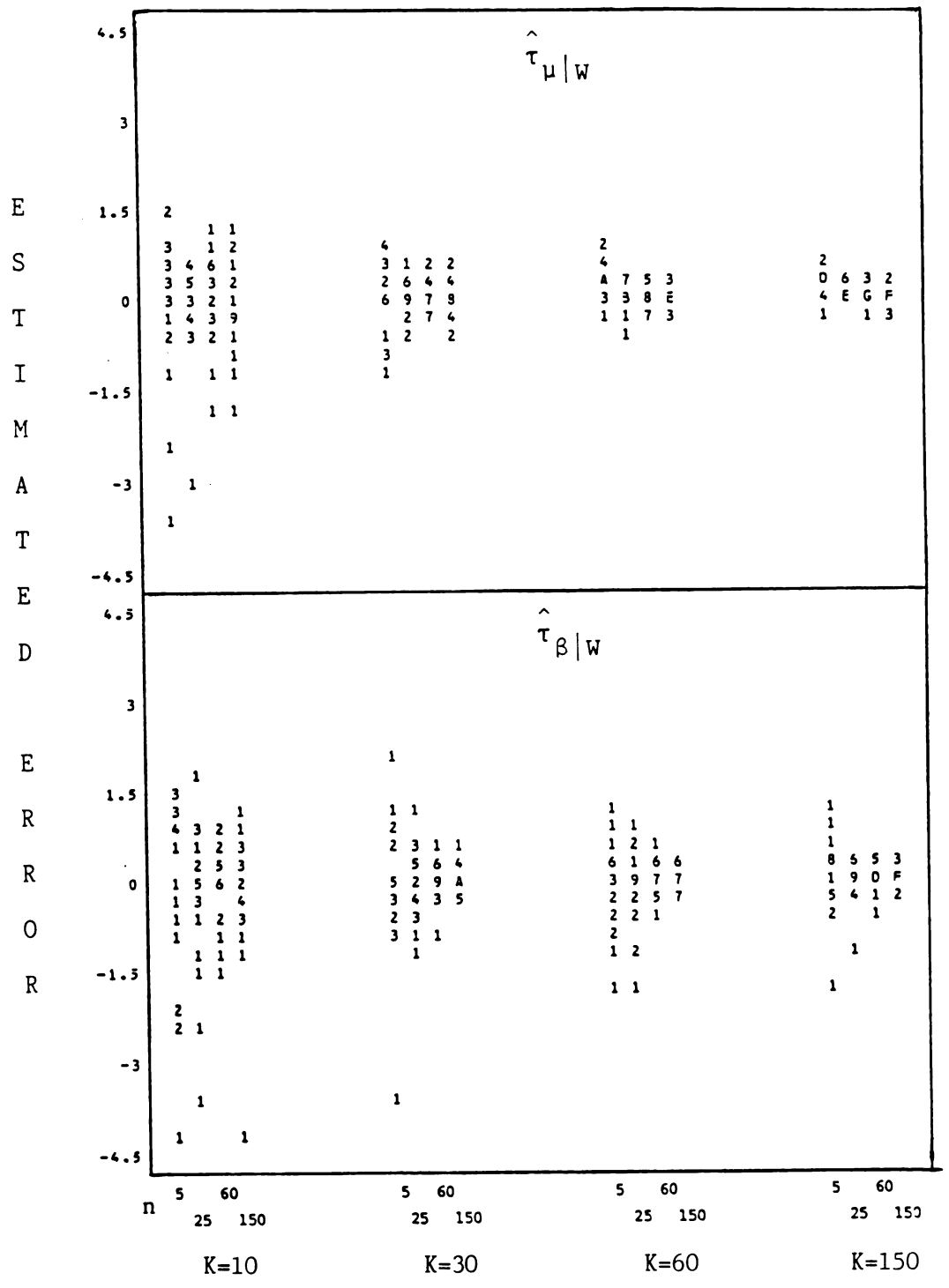

Figure A-3. Error estimates in transformed $\hat{\tau}_{\mu|W}$ and $\hat{\tau}_{\beta|W}$ for different combinations of k and n with d=.25.

Note: Symbols A-z signify frequencies 10-35, respectively.

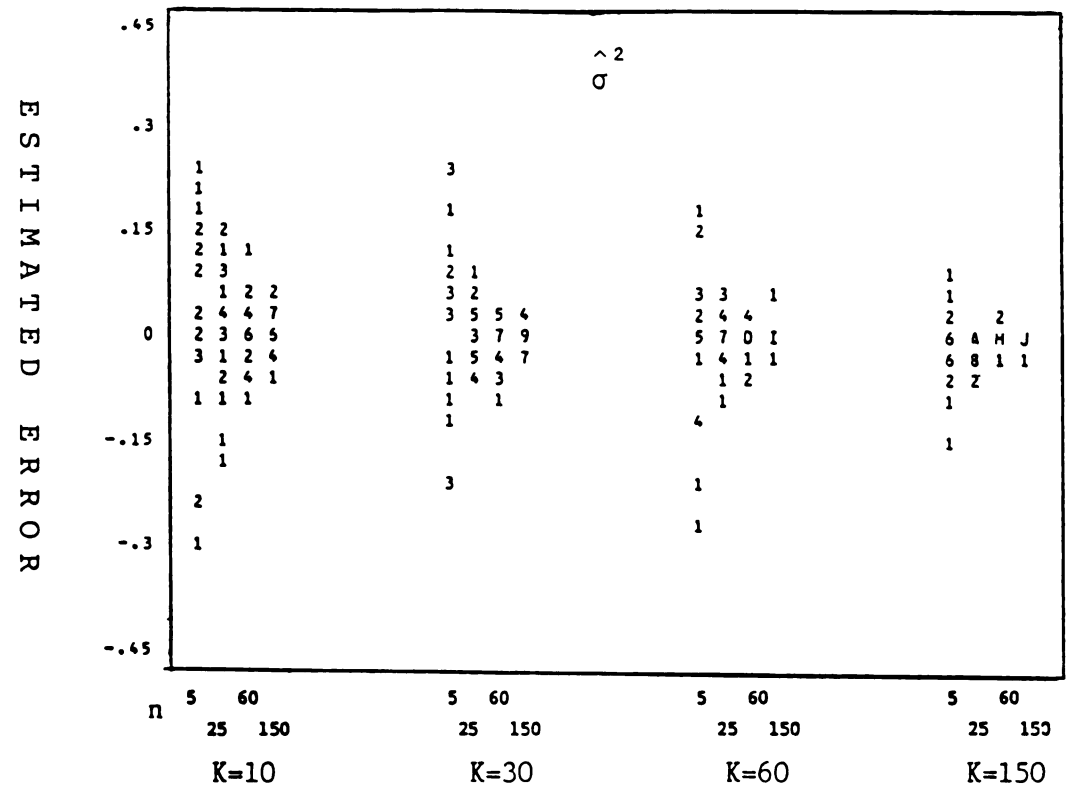

Figure A-4. Error estimates in transformed $\hat{\sigma}^2$ for different combinations of k and n with d=.25.

Note: Symbols A-Z signify frequencies 10-35, respectively.

Table A-1

Mean and Variance of the Z-Statistics for the Macro Parameters for Different Number of Groups*

K		$Z_{\gamma_{00}}$	$Z_{\gamma_{01}}$	$Z_{\gamma_{10}}$	$Z_{\gamma_{11}}$
10	Mean	.004	.085	.005	.073
	Variance	1.185	1.187	.956	1.079
30	Mean	.070	-.087	-.104	-.007
	Variance	1.120	1.094	1.248	1.232
60	Mean	.048	.001	.039	.066
	Variance	.847	1.101	1.018	1.008
150	Mean	.031	-.001	-.102	.158
	Variance	1.153	1.171	.971	1.195

*From 160 replications.

Table A-2

Mean and Variance of the Z-Statistics for the Macro Parameters for Different Group Sizes*

n		$Z_{\gamma_{00}}$	$Z_{\gamma_{01}}$	$Z_{\gamma_{10}}$	$Z_{\gamma_{11}}$
5	Mean	.111	.034	.004	.161
	Variance	1.084	1.292	.990	.914
25	Mean	.080	-.007	-.071	-.007
	Variance	1.023	.943	1.130	1.212
60	Mean	.009	-.034	.016	.028
	Variance	1.018	1.235	1.145	1.105
150	Mean	-.048	.005	-.111	.107
	Variance	1.167	1.095	.934	1.279

*From 160 replications.

Table A-3

Mean and Variance of the Z-Statistics for the Macro parameters for Different Intraclass Correlation Coefficients*

d		$Z_{\gamma_{00}}$	$Z_{\gamma_{01}}$	$Z_{\gamma_{10}}$	$Z_{\gamma_{11}}$
.10	Mean	.116	.052	-.056	.043
	Variance	.970	1.062	1.080	1.192
.25	Mean	-.040	-.053	-.025	.102
	Variance	1.165	1.209	1.018	1.063

*From 320 replications.

Table A-4

Power for Tests of Macro Parameter γ_{01}^{*}

			$\alpha = .01$		$\alpha = .05$		$\alpha = .10$	
k	n	e.s.	p'	p''	p'	p''	p'**	p'''***
10	5	.31	.1423	.1762	.3264	.3783	.4443	.5000
	25	.32	.2005	.1762	.4090	.3783	.5359	.5000
	60	.31	.2514	.2327	.4801	.4562	.6026	.5793
	150	.32	.4761	.2709	.7088	.5040	.8078	.6255
30	5	.31	.3336	.4483	.5714	.6844	.6915	.7881
	25	.32	.7486	.7517	.9015	.9032	.9452	.9474
	60	.31	.7517	.7454	.9032	.8997	.9463	.9441
	150	.32	.8106	.8315	.9332	.9429	.9656	.9706
60	5	.31	.6331	.7823	.8315	.9192	.8980	.9564
	25	.32	.9783	.9854	.9957	.9974	.9984	.9990
	60	.31	.9940	.9934	.9990	.9990	.9997	.9997
	150	.32	.9951	.9946	.9992	.9991	.9998	.9997
150	5	.31	.9898	.9987	.9984	.9998	.9994	.9999
	25	.32	.9999	.9999	.9999	.9999	.9999	.9999
	60	.31	.9999	.9999	.9999	.9999	.9999	.9999
	150	.32	.9999	.9999	.9999	.9999	.9999	.9999

*From 40 replications of k groups of size n with e.s. effect size.
**Empirical power
***Nominal power

Table A-5

Power for Tests of Macro Parameter

γ_{10}*

			α =.01		α =.05		α =.10	
k	n	e.s.	p'	p"	p'	p"	p'**	p"***
10	5	.73	.8051	.8980	.9306	.9699	.9633	.9861
	25	.78	.9999	.9999	.9999	.9999	.9999	.9999
	60	.73	.9999	.9999	.9999	.9999	.9999	.9999
	150	.78	.9999	.9999	.9999	.9999	.9999	.9999
30	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
	25	.78	.9999	.9999	.9999	.9999	.9999	.9999
	60	.73	.9999	.9999	.9999	.9999	.9999	.9999
	150	.78	.9999	.9999	.9999	.9999	.9999	.9999
60	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
	25	.78	.9999	.9999	.9999	.9999	.9999	.9999
	60	.73	.9999	.9999	.9999	.9999	.9999	.9999
	150	.78	.9999	.9999	.9999	.9999	.9999	.9999
150	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
	25	.78	.9999	.9999	.9999	.9999	.9999	.9999
	60	.73	.9999	.9999	.9999	.9999	.9999	.9999
	150	.78	.9999	.9999	.9999	.9999	.9999	.9999

*From 40 replications of k groups of size n with e.s. effect size.
**Empirical power
***Nominal power

Table A-6

Power for Tests of Macro Parameter

γ_{11}*

			α =.01		α =.05		α =.10	
k	n	e.s.	p'	p"	p'	p"	p'**	p"***
10	5	.16	.0418	.0495	.1314	.1515	.2119	.2358
	25	.17	.1112	.1170	.2709	.2843	.3859	.3974
	60	.16	.2451	.2061	.4681	.4207	.5948	.5478
	150	.17	.2676	.2005	.5000	.4129	.6217	.5359
30	5	.16	.1492	.1788	.3372	.3783	.4562	.5040
	25	.17	.3974	.3669	.6406	.6103	.7486	.7224
	60	.16	.5160	.5080	.7454	.7357	.8340	.8289
	150	.17	.9015	.8665	.9719	.9573	.9868	.9793
60	5	.16	.2546	.2810	.4840	.5160	.6064	.6368
	25	.17	.8577	.8531	.9535	.9515	.9772	.9761
	60	.16	.9656	.9573	.9925	.9904	.9970	.9961
	150	.17	.9871	.9854	.9978	.9974	.9991	.9990
150	5	.16	.8461	.8888	.9484	.9664	.9744	.9842
	25	.17	.9987	.9987	.9999	.9998	.9999	.9999
	60	.16	.9999	.9999	.9999	.9999	.9999	.9999
	150	.17	.9999	.9999	.9999	.9999	.9999	.9999

*From 40 replications of k groups of size n with e.s. effect size.
**Empirical power
***Nominal power

Table A-7

Power for Tests of Macro Parameter

γ_{01}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p"	p'	p"	p'**	p"***
.10	10	5	.22	.0735	.1251	.2005	.2981	.3015	.5871
		25		.0838	.0778	.2206	.2119	.3264	.3121
		60		.1841	.1587	.3859	.3520	.5120	.4721
		150		.4681	.2776	.7054	.4920	.8023	.6331
	30	5		.2611	.4207	.4880	.6591	.6141	.7673
		25		.6950	.7019	.8708	.8729	.9251	.9279
		60		.6808	.6736	.8599	.8577	.9192	.9162
		150		.8133	.8485	.9332	.9505	.9656	.9750
	60	5		.5557	.7257	.7734	.8888	.8577	.9370
		25		.9778	.9875	.9957	.9978	.9984	.9991
		60		.9927	.9934	.9988	.9990	.9996	.9997
		150		.9920	.9927	.9987	.9988	.9995	.9996
	150	5		.9319	.9875	.9826	.9978	.9922	.9991
		25		.9999	.9999	.9999	.9999	.9999	.9999
		60		.9999	.9999	.9999	.9999	.9999	.9999
		150		.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

Table A-8

Power for Tests of Macro Parameter
γ_{01}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p''	p'	p''	p'**	p'''**
.25	10	5	.39	.2451	.2389	.4721	.4641	.5948	.5871
		25	.43	.3783	.3336	.6217	.5714	.7324	.6915
		60	.39	.3300	.3228	.5714	.5636	.6915	.6808
		150	.43	.4801	.2676	.7157	.5000	.8106	.6217
	30	5	.39	.4129	.4721	.6554	.7088	.7611	.8051
		25	.43	.7967	.7995	.9265	.9279	.9608	.9625
		60	.39	.8159	.8078	.9345	.9319	.9664	.9641
		150	.43	.8106	.8133	.9332	.9345	.9649	.9656
	60	5	.39	.7054	.8315	.8770	.9418	.9292	.9706
		25	.43	.9783	.9834	.9957	.9970	.9984	.9988
		60	.39	.9952	.9934	.9993	.9990	.9997	.9997
		150	.43	.9969	.9960	.9996	.9994	.9999	.9998
	150	5	.39	.9991	.9999	.9999	.9999	.9999	.9999
		25	.43	.9999	.9999	.9999	.9999	.9999	.9999
		60	.39	.9999	.9999	.9999	.9999	.9999	.9999
		150	.43	.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

Table A-9

Power for Tests of Macro Parameter
γ_{10}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p"	p'	p"	p'**	p"***
.10	10	5	.74	.7224	.8599	.8849	.9554	.9357	.9783
		25	.71	.9999	.9999	.9999	.9999	.9999	.9999
		60	.74	.9999	.9999	.9999	.9999	.9999	.9999
		150	.71	.9999	.9999	.9999	.9999	.9999	.9999
	30	5	.74	.9999	.9999	.9999	.9999	.9999	.9999
		25	.71	.9999	.9999	.9999	.9999	.9999	.9999
		60	.74	.9999	.9999	.9999	.9999	.9999	.9999
		150	.71	.9999	.9999	.9999	.9999	.9999	.9999
	60	5	.74	.9999	.9999	.9999	.9999	.9999	.9999
		25	.71	.9999	.9999	.9999	.9999	.9999	.9999
		60	.74	.9999	.9999	.9999	.9999	.9999	.9999
		150	.71	.9999	.9999	.9999	.9999	.9999	.9999
	150	5	.74	.9999	.9999	.9999	.9999	.9999	.9999
		25	.71	.9999	.9999	.9999	.9999	.9999	.9999
		60	.74	.9999	.9999	.9999	.9999	.9999	.9999
		150	.71	.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

Table A-10

Power for Tests of Macro Parameter
γ_{10}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p"	p'	p"	p'**	p"***
.25	10	5	.73	.8729	.9265	.9599	.9808	.9808	.9913
		25	.85	.9890	.9854	.9982	.9974	.9993	.9990
		60	.73	.9999	.9999	.9999	.9999	.9999	.9999
		150	.85	.9999	.9990	.9999	.9999	.9999	.9999
	30	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
		25	.85	.9999	.9999	.9999	.9999	.9999	.9999
		60	.73	.9999	.9999	.9999	.9999	.9999	.9999
		150	.85	.9999	.9999	.9999	.9999	.9999	.9999
	60	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
		25	.85	.9999	.9999	.9999	.9999	.9999	.9999
		60	.73	.9999	.9999	.9999	.9999	.9999	.9999
		150	.85	.9999	.9999	.9999	.9999	.9999	.9999
	150	5	.73	.9999	.9999	.9999	.9999	.9999	.9999
		25	.85	.9999	.9999	.9999	.9999	.9999	.9999
		60	.73	.9999	.9999	.9999	.9999	.9999	.9999
		150	.85	.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

Table A-11

Power for Tests of Macro Parameter γ_{11}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p"	p'	p"	p'**	p"***
.10	10	5	.12	.0228	.0294	.0838	.1003	.1446	.1685
		25	.10	.0694	.0823	.1922	.2177	.2912	.3228
		60	.12	.2005	.1446	.4129	.3300	.5359	.4483
		150	.10	.3156	.1841	.5517	.3859	.6736	.4880
	30	5	.12	.1112	.1469	.2743	.3336	.3859	.4522
		25	.10	.3015	.2676	.5359	.4960	.6591	.6217
		60	.12	.2643	.2578	.4920	.4840	.6179	.6103
		150	.10	.8212	.7734	.9382	.9147	.9678	.9535
	60	5	.12	.0934	.1093	.2420	.2709	.3520	.3859
		25	.10	.8389	.8365	.9463	.9452	.9726	.9719
		60	.12	.9082	.8869	.9744	.9664	.9881	.9990
		150	.10	.9812	.9788	.9965	.9960	.9987	.9985
	150	5	.12	.7088	.7881	.8790	.9207	.9306	.9582
		25	.10	.9949	.9934	.9991	.9989	.9997	.9997
		60	.12	.9999	.9998	.9999	.9999	.9999	.9999
		150	.10	.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

Table A-12

Power for Tests of Macro Parameter
γ_{11}*

				α =.01		α =.05		α =.10	
d	k	n	e.s.	p'	p"	p'	p"	p'**	p"***
.25	10	5	.19	.0708	.0808	.1977	.2148	.2946	.3192
		25	.24	.1685	.1611	.3632	.3557	.4880	.4801
		60	.19	.2912	.2843	.5279	.5160	.6480	.6406
		150	.24	.2266	.2177	.4483	.4364	.5714	.5596
	30	5	.19	.1977	.2119	.4052	.4247	.5319	.5517
		25	.24	.5000	.4801	.7324	.7123	.8238	.8106
		60	.19	.7642	.7549	.9099	.9032	.9505	.9474
		150	.24	.9515	.9279	.9887	.9808	.9953	.9916
	60	5	.19	.4960	.5239	.7291	.7517	.8212	.8413
		25	.24	.8729	.8665	.9608	.9582	.9808	.9793
		60	.19	.9893	.9875	.9982	.9979	.9993	.9991
		150	.24	.9916	.9904	.9987	.9984	.9995	.9994
	150	5	.19	.9306	.9495	.9821	.9878	.9920	.9949
		25	.24	.9997	.9998	.9999	.9999	.9999	.9999
		60	.19	.9999	.9999	.9999	.9999	.9999	.9999
		150	.24	.9999	.9999	.9999	.9999	.9999	.9999

*From 20 replications of k groups of size n with d intraclass correlation coefficient of e.s. effect size.
**Empirical power
***Nominal power

REFERENCES

References

Aitkin, M., Anderson, D. and Hinde, J. (1981) "Statistical Modeling of Data on Teaching Styles", Journal of the Royal Statistical Society, Series A, 144(4), 419-461.

Aitkin, M., and Longford, N. (1986) "Statistical Modeling Issues in School Effectiveness Studies", Journal of the Royal Statistical Society, Series A, 149(1), 1-43.

Alker, H.R. (1969) "A Typology of Ecological Fallacies", in M. Dogan and S. Rokkan (eds.), Social Ecology. Cambridge, MA: MIT Press.

Anderson, T.W. (1969) "Statistical Inference for Covariance Matrices with Linear Strucutre", in P.R. Krishnaiah (ed.), Multivariate Analysis-II, 55-66. New York: Academic Press.

------. (1971) Estimation of Covariance Matrices with Linear Structure and Moving Average Processes of Finite Order. Technical Report No. 6, Department of Statistics, Stanford University, Stanford, CA.

Bartlett, M.S., and Kendall, D.G. (1946) "The Statistical Analysis of Variances- Heterogeneity and the Logarithmic Transformation", Journal of the Royal Statistical Society, 8, 128-138.

Bock, R.D. (1983) "The Discrete Bayesian", in H. Wainer and Messick (eds.), Principles of Modern Psychological Measurement: A Festschrift for Fredrick M. Lord. Hillsdale, N.J.: Lawrence Erlebaum Associates.

Boyd, L.H. Jr., and Iversen, G.R. (1979) Contextual Analysis: Concepts and Statistical Techniques. Belmont, CA: Wadsworth.

Braun, H.I.; Jones, D.H.; Robin, D.B.; and Thayer, D.T. (1983) "Empirical Bayes Estimation of Coefficients in the General Linear Model With Data of Deficient Rank", Psychometrika, 48(2), 171-181.

Bryk, A.S., and Raudenbush, S.W. (1987) "Application of Hierarchical Linear Models to Assessing Change", Psychological Bulletin, 101(1), 147-158.

Bryk, A.S., Raudenbush, S.W.; Seltzer, M.; and Congdon (1987) An Introduction to HLM: Computer Program and User's Guide. Version 1.0, University of Chicago.

Burstein, L. (1980) "The Analysis of Multi-Level Data in Educational Research and Evaluation", Review of Research in Education, 8, 158-233.

Burstein, L.; Linn, R.L.; and Capell, F. (1978) "Analyzing Multilevel Data in the Presence of Heterogeneous Within-class Regression", Journal of Educational Statistics, 3(4), 347-383.

Burstein, L.; Miller, M.D.; and Linn, R.L. (1979) The use of Within-Group Slopes as Indices of Group outcomes. Unpublished Manuscript.

Burstein, L. and Miller, M.D. (1980) "Regression-Based Analysis of Multi-Level Educational Data", New Directions for Methodology of Social and Behavioral sciences, 6, 194-211.

Cooley, W.W.; Bond, L.; and Mao, B. (1981) "Analyzing Multi-Level Data", in R.A. Berk (ed.), Educational Evaluation Methodology. Baltimore, MD: John Hopkins University Press.

Cronbach, L.J. (1976) Research on Classrooms and Schools: Formulation of questions, Design and Analysis. Stanford, CA: Standard Evaluation Consortium.

Cronbach, L.J., and Webb, N. (1975) "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, 67, 717-724.

de Leeuw, J.D., and Kreft,I. (1986) "Random Coefficients Models for Multilevel Analysis", Journal of Educational Studies, 11, 57-85.

Dempster, A.P.; Laird, N.M.; and Rubin, D.B. (1977) "Maximum Likelihood From Incomplete Data via the EM Algorithm (with discussion)", Journal of the Royal Statistical Society, Series B, 39, 1-18.

Dempster, A.P.; Rubin, D.B.; and Tsutakawa, R.D. (1981) "Estimation in Covariance Components Models", Journal of American statistical Association, 76, 341-353.

DerSimonian, R., and Laird, N.M. (1983) "Evaluating the Effect of Coaching on SAT Scores: A Meta-Analysis", Harvard Educational Review, 53(1), 1-15.

Dielman, T.E. (1983) "Pooled Cross-Sectional and Time-series Data: A Survey of Current Statistical Methodology", The American Statistician, 37, 111-122.

Efron, B. and Morris, C. (1975) "Data analysis Using Stein's Estimator and its Generalizations", Journal of the American Statistical Association, 74, 311-319.

Elston, R.C., and Grizzle, J.E. (1962) "Estimation of Time-Response Curves and Their Confidence Bands", Biomtrics, 18, 148-159.

Goldstein, H.(1986) "Multilevel Mixed Linear Model analysis Using Iterative Generalized Least Squares", Biometrika, 73(1), 43-56.

Graybill, F.A. (1976) Theory and Application of the linear Model. Belmont, CA: Wadsworth Publishing Co.

Hammersley, J.M., and Handscomb, D.C. (1964) Monte Carlo Methods. New York: Barnes and Noble, Inc.

Haney, W. (1980) "Units and Levels of Analysis in Large-Scale Evaluation", New Directions for Methodology of Social and Behavioral Sciences, 6, 1-15.

Hannan, M.T. (1971) Aggregation and Disaggregation in Sociology. Lexington, MA: D.C. Heath.

Hartley, H.O., and Rao, J.N.K. (1967) "Maximum Likelihood Estimation for the Mixed Analysis of Variance Model", Biometrika, 54, 93-108.

Harville, D.A. (1977) "Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems", Journal of American Statistical Association, 72, 320-340.

Henderson, C.R. (1953) "Estimation of Variance and Covariance Components", Biometrics, 9, 226-52.

Hocking, R.R., and Kutner, M.H. (1975) "Some Analytical and Numerical Comparisons of estimators for the Mixed ANOVA Model", Biometrics, 31, 19-28.

Kackar, R.N., and Harville, D.A. (1984) "Approximations for Standard error of Estimation of fixed and Random Effects in Mixed Linear Models", Journal of the American Statistical Association, 79, 388, 853-862.

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

Keesling, J.W., and Wiley, D.E. (1974) Regression Models for Hierarchical Data . Paper presented at the Annual Meeting of the Psychometrics Society.

Kirk, R.E. (1968) Experimental Design: Procedures for the Behavioral Sciences. Belmont, CA: Brooks/Cole Publishing co.

Kish, L. (1983) Survey Sampling. Malabar, Fla.: R.E. Krieger.

Klotz, J.H.; Milton, R.C.; and Zacks, S. (1969) "Mean Square Efficiency of Estimators of Variance Components", Journal of the American Statistical Association, 64, 1383-1402.

Knapp, T.R. (1977) "The Unit of Analysis Problem in Applications of Simple Correctional Research", Journal of Educational Statistics, 2(3), 171-186.

Knuth, D. (1969) Seminumerical Algorithms. Reading, Mass.: Addison-Wesley.

Laird, N.M., and Ware, J.H. (1982) "Random-Effects Models for Longitudinal Data", Biometrics, 38, 963-974.

Lee, V.E. (1986) Multi-Level Causal Models for Social Class and Achievement. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Linley, D.V., and Smith, A.F.M. (1972) "Bayes Estimates for the Linear Model", Journal of the Royal Statistical Society, Series B, 34,1-41.

Longford, N.T. (1985) A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Effects. An unpublished paper, Institute of Applied Statistics. Lancaster University, Lancaster, England.

Mason, W.M.; Wong, G.Y.; and Entwisle, B. (1984) "Contextual Analysis Through the Multi-level Linear Model", Sociological Methodology,. San Fransisco, CA: Jossey Bass, 72-103.

Miller, J. (1973) Asymptotic Properties and Computation of Maximum Likelihood Estimates in the Mixed Model of the Analysis of Variance. Technical Report No. 12, Department of Statistics. Stanford University, Stanford, California.

------. (1977) "Asymptotic Properties of Maximum Likelihood Estimates in the Mixed Model of Analysis of Variance",

Annals of Statistics, 5, 746-762.

Morris, C.N. (1983) "Parametric Empirical Bayes Inference: Theory and Applications", Journal of the American Statistical Association, 78(381), 47-65.

Novick, M.R.; Jackson, P.H.; Thayer, D.T.; and Cole, N.S. (1972) "Estimating Multiple Regressions in m Groups- A Cross-Validation Study", British Journal of Mathematical and Statistical Psychology, 25, 33-50.

Novick, M.R., and Jackson, P.H. (1974) "Further Cross-Validation analysis of the Bayesian m-Groups Regression Method", American Educational Research Journal , 11, 77-85.

Olsen, A; Seely, J.; and Brikes, D. (1976) "Invariant Quadratic Unbiased Estimation for Two Variance Components", Annals of Statistics, 4, 878-890.

Patterson, H.D., and Thompson, R. (1974) Maximum Likelihood Estimation of Components of Variance, Proceedings of the 8th International Biometric Conference, 197-207.

Patterson, H.D., Thompson, R. (1971) "Recovery of Inter-Block Information when Block Sizes are Unequal", Biometrika, 58, 545-554.

Pedhazur, E.J. (1973) Multiple Regression in Behavioral Research. New York: Holt, Reinhart and Winston.

Pitman, E.J.G. (1938) "The Estimation of the Location and Scale Parameters of a Continuous Population of any Given Form", Biometrika, 30, 391-421.

Rao, C.R. (1972) "Estimation of Variance and Covariance Components in linear Models", Journal of the American Statistical association, 67, 112-115.

Raudenbush, S.W. (1984) Application of a Hierarchical Linear Model in educational Research . Unpublished Doctoral Dissertation. Harvard University: Graduate School of Education.

Raudenbush, S.W., and Bryk, A.S. (1986) "Empirical Bayes Meta-Analysis", Journal of Educational statistics, 10, 75-98.

Raudenbush, S.W., and Bryk, A.S. (1986) "A Hierarchical Model for Studying School Effects", Sociology of Education, 59, 1-17.

Raudenbush, S.W. (1988) "Educational Applications of Hierarchical Linear Models: A Review", Journal of Educational

Satistics, 13(2), 85-116.

Robinson, W.S. (1950) "Ecological Correlations and the Behavior of Individuals", American Sociological Review, 15, 351-357.

Rogosa, D. (1978) "Politics, Process, and Pyramids", Journal of Educational Statistics, 3(1), 79-86.

Rosenberg, B. (1973) "Linear Regression with Randomly Dispersed Parameters", Biometrika, 60, 61-75.

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

-------. (1981) "Estimation in Parallel Randomized Experiments", Journal of Educational Statistics, 6(4), 377-400.

------. (1983) "Some Applications of Bayesian Statistics to Educational data," The Statistician, 32, 55-68.

Schmidt, W.H. (1969) Covariance Structure Analysis of the Multi-variate Random Effects Model. Unpublished dissertation, University of Chicago.

Schmidt, W.H., and Houang, R.T. (1983) A Comparison of Three Analytical Strategies for Hierarchical Data. Paper presented at the Annual Meeting of the American Educational Research Association, Los Angeles, California.

Searle, S.R. (1968) "Another Look at Henderson's Methods of Estimating Variance Components", Biometrics, 24, 749-788.

------. (1971a) Linear Models. New York: Wiley & Sons.

------. (1971b) "Topics in Variance Component Estimation", Biometrics, 27,1-76.

Seely, J. (1975) "An Example of an Inquadmissible Analysis of Variance Estimator for a Variance Component", Biometrika, 62, 689-90.

Shigemasu, K. (1976) "Development and Validation of a Simplified m-Group Regression Model", Journal of Educational Statistics, 1(2), 157-180.

Smith, A.F.M. (1973) "A General Bayesian Linear Model", Journal of the Royal Statistical Society, Series B, 35, 67-75.

Sternio, J.L.F. (1981) Empirical Bayes Estimation for a

Hierarchical Linear Model . Unpublished dissertation, Department of Statistics, Harvard University.

Sternio, J.L.F.; Weisberg, H.I.; and Bryk, A.S. (1983) "Empirical Bayes Estimation of Individual Growth Curves Parameters and Their Relationship to Covariates", Biometrics, 39, 71-86.

Swamy, P.A.V. (1973) "Criteria, Costraints, and Multicollinearity in Random Coefficient Regression Models", Annals of Economic and Social Measurement, 4, 429-450.

Tate, R.L., and Wongbundhit, Y. (1983) "Random Versus Nonrandom Coefficient Model for Multilevel Analysis", Journal of Educational Statistics, 8, 103-120.

Walsh, J.E. (1947) "Concerning the effect of the Intraclass Correlation on Certain Significance Tests", Annals of Mathematical Statistics, 18, 88-96.

Wiley, D.E. (1970) "Design and Analysis of Evaluation Studies", in Witrock, M.C., and Wiley, D.E. (eds.), The Evaluation of Instruction. New York: Holt, Rinehart and Winston.

------. (1976) "Another Hour, Another Day: Quality of Schooling a Potent Path for Policy", in Sewell, W.H.; Hauser, R.M.; and Featherman, D.L. (eds.), Schooling and Achievement in American Society, New York: Academic.

Wisenbacker, J., and Schmidt, W.H. (1979) The Structural Analysis of Hierarchical Data. Paper presented at the annual meeting of the American Educational Research Association.

Wong, G.Y., and Mason, W.M. (1985) "The Hierarchical Logistic Regression Model for Multilevel Analysis", Journal of American Statistical Association, 80, 513-524.