This is to certify that the

dissertation entitled

A COMPARISON OF THE EM AND DATA AUGMENTATION ALGORITHMS
ON SIMULATED SMALL SAMPLE HIERARCHICAL DATA
FROM RESEARCH ON EDUCATION

presented by

Randall Peter Fotiu

has been accepted towards fulfillment
of the requirements for

Ph.D degree in Educational Psychology and Special Education

Major professor

Date February 17, 1989

A COMPARISON OF THE EM AND DATA AUGMENTATION ALGORITHMS ON SIMULATED SMALL SAMPLE HIERARCHICAL DATA FROM RESEARCH ON EDUCATION

By

Randall Peter Fotiu

AN ABSTRACT OF 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

1989

ABSTRACT

A COMPARISON OF THE EM AND DATA AUGMENTATION ALGORITHMS ON SIMULATED SMALL SAMPLE HIERARCHICAL DATA FROM RESEARCH ON EDUCATION

By

Randall Peter Fotiu

The hierarchical organization of our schools is reflected in the data collected in a natural school setting by educational researchers. It is common for researchers investigating the effects of an experimental treatment at the class level to obtain data from a small number of classes and an unequal number of students within each class. The statistical analysis of this type of data provides many challenges, especially when there is an interaction between a student level characteristic and the class level treatment. The parameters and notation of the specific hierarchical linear model are developed in detail. Two statistical procedures are evaluated on simulated small sample hierarchical data. The empirical Bayes procedure using the EM algorithm is compared to the Bayesian procedure implemented with the data augmentation algorithm. Detailed explanations are provided that pertain to the application of these two statistical procedures and the algorithms used in their respective implementations.

The data used to evaluate these two statistical procedures were simulated to represent a modest experimental design. This design included a measured student

outcome and an independent variable representing a student characteristic such as motivation, aptitude, or interest. Furthermore, students were nested within ten classes and these classes were randomly assigned to either a treatment or control group. The size of each class was randomly selected from a distribution of class size. This data was generated to reflect parameter values of the hierarchical linear model that might typically be obtained. The interaction effect between student and class level factors was evaluated at three effect sizes. For each level of the interaction effect, five hundred replications of an experiment were generated.

The Bayesian approach using the data augmentation algorithm recovered the parameters of the hierarchical linear model as well or better than the empirical Bayes approach using the EM algorithm. In particular, the Bayesian approach was superior in recovering the between-class variance-covariance parameter matrix. In addition, the Bayesian approach using the data augmentation algorithm provided finite approximations to the posterior distributions of the parameters of the model. The empirical Bayes approach using the EM algorithm performed better in hypothesis testing of the between class effect parameters when a t-test was utilized. The empirical Bayes procedure using a z-test and this implementation of the Bayesian procedure tended to provide liberal hypothesis tests. The implementation of the Bayesian procedure using the data augmentation algorithm required considerable computer resources, whereas the implementation of the empirical Bayes procedure using the EM algorithm was well within practical limits for research on education.

ACKNOWLEDGMENTS

I would like to thank my dissertation director Stephen Raudenbush and dissertation committee members Andrew Porter, Richard Houang, Susan Melnick and James Stapleton, for their advice, insight and guidance throughout the process of writing this dissertation. I have been very fortunate to have had the opportunity to work with this outstanding group of researchers and teachers throughout my graduate studies. My wife, Josie Wojtowicz, deserves special thanks for her support and understanding while I wrote this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

AN OVERVIEW OF THE STRUCTURE AND ANALYSIS OF HIERARCHICAL DATA IN EDUCATION

Introduction

Throughout history, scientific inquiry on a specific subject has often raised interesting questions in unexpected areas. Not only has science taken unexpected turns, it has fostered the continual improvement and advancement of its common methods and tools. Progress in statistics and experimental design has moved forward hand-in-hand with science. William Sealy Gosset was hired by the Arthur Guinness Sons & Company, Limited of Dublin in 1899 to help apply the scientific method to the brewing process. During his work at the brewery, it became evident that formal statistical analysis needed to be extended to small samples. Frequently, samples of eight to twelve were obtained at the brewery. This supplied Gosset with the motivation to develop the "Student's t-test" to provide a statistical tool for small sample problems (Tankard, 1984). Sir Ronald Fisher's work at the Rothamsted Experimental Station provided a substantive context for his work in the development of factorial designs and the analysis of variance method. In the preface of Fisher's classic work, *Statistical Methods for Research Workers* (1936, p. vii), he stated, "Daily contact with the statistical problems which present themselves to the laboratory worker has stimulated the purely mathematical researches upon which are based the methods here presented." Research into the educational process as it

takes place in schools also provides a unique environment with many associated methodological challenges.

Much of the research interest in education is focused on the effects of various experimental treatments applied in a natural school setting. Not only are the main effects directly attributable to the experimental treatments of interest, the interaction effects these treatments have with known student characteristics and attributes are also critical concerns of educators. This combination of experimental control of treatments, the natural school settings in which they are applied, the psychological characteristics of students, and the financial constraints of most educational research present many difficulties for statisticians involved in this area of research. The issues related to the social organization of schools and the structure of experimental data found in educational research are developed in the following sections of this chapter. Many statistical procedures have been formulated to cope with the problems presented by data generated in this type of research setting. A critical review of some of the most widespread statistical procedures applied in this research setting is also presented.

The historical advancement of statistical methods used in experimental research in education leads naturally to a Bayesian formulation. The new data augmentation algorithm developed by Tanner and Wong (1987) shows promise as a practical implementation of a Bayesian solution and resolves many of the shortcomings found in prior statistical methods used in educational research. To illustrate its application in an educational research context, this new algorithm is implemented in a computer program and applied to simulated data sets that mimic the characteristics of educational data. And to evaluate the performance of this new

algorithm, it is compared to an alternative statistical estimation procedure, the empirical Bayes approach using the EM Algorithm developed by Dempster, Laird and Rubin (1977).

The Structure of Educational Data and Associated Problems

The majority of educational research takes place in a school setting. As a consequence, the hierarchical organization of our schools is reflected in the data collected in a natural school setting. The statistical analysis of this hierarchical data provides many challenges. This hierarchical data structure resulting from research in schools, the lengthy debate amongst educational researchers concerning the appropriate unit of analysis, the implications of unbalanced designs, and small sample problems are methodological issues encountered in research on education. The following sections address these important and related issues.

The Hierarchical Organization of Educational Data

Typically, the prominent interest of educational research is to investigate how characteristics of the school or classroom environment influence student learning (Good and Brophy, 1986; and Brophy and Good, 1986). In some cases, the characteristics of interest are an inherent part of the existing setting, and in others they are introduced and controlled by the experimenter. In educational studies, both naturally occurring and experimentally controlled factors are commonly found together. Since a primary goal of educational research is to apply promising positive results in an effort to improve student learning, it makes sense to study the educational process in a natural setting. On one hand, an inference is the most

generalizable when the research and the application settings are equivalent or closely approximate one another. However, conducting research in a natural setting limits experimental control.

A feature of our formal education system and its social organization is its hierarchical (nested) structure. For example, students are grouped within classrooms, classrooms are grouped within schools, and schools are grouped within districts. This multilevel structure provides a conceptual framework for understanding the effects of schooling (Bidwell and Kasarda, 1980; and Barr and Dreeden, 1983). The resources available to a teacher, such as books, materials, and class size, have an influence on the instruction provided to students. The same instruction provided to students may have differential effects (Cronbach and Snow, 1977) and this differential response may cause the teacher to update and adjust their lesson plans. Thus, we have a model for schooling that is both hierarchical in structure and the levels of the hierarchy are interdependent. This conceptual model of schooling should be reflected in the statistical models used to evaluate and understand the effects of schooling (Burstein, 1980; Cooley, Bond and Mao, 1981; Raudenbush and Bryk, 1986; Goldstein, 1987).

The Unit of Analysis Debate and the Specification of the Hierarchical Model

This multilevel and interactive model of schooling has prompted a debate among researchers on the appropriate unit of analysis. Should educational research at the class level be evaluated strictly at the class level, the student level, or some combination? In addition, how can this model be specified to reflect the complexities of the hierarchical data collected from educational research? These

questions are discussed in this section.

To exemplify the unit of analysis problem, let us construct a hypothetical research study. Suppose a sample of ten classes were drawn randomly from the population of interest and for illustrative purposes, each class contained 20 students. Five of the classes are randomly assigned to an experimental condition and the remaining classes are assigned to a control group. Furthermore, the experimental group's treatment is applied at the class level and differs from the control group only in the sequence in which the material is presented. The research question of interest is whether the new experimental sequencing of the educational material promotes higher levels of student achievement when compared to the existing order of presentation. When the unit of analysis is the individual student, then an independent group t-test with 198 degrees of freedom (practically equivalent to a z-test) could be used to test for the effect of the experimental condition relative to the control condition. On the other hand, if the class is taken as the unit of analysis, a t-test with 8 degrees of freedom could be used to test for a treatment effect. Under the assumptions that the observations are independent and drawn from the same normally distributed population, the independent group t-test using student level data is substantially more powerful than the alternative t-test using class level data. Hence, for a researcher under pressure to find significant differences, the choice is clearly in favor of the more powerful t-test using student level data. But, are the assumptions for the independent t-test satisfied when comparing treatments that are presented at the class level?

The assumption that individual students within a class are statistically independent may not be appropriate. Glass and Stanley (1970, pp. 505-506) make

the distinction between the "units of statistical analysis" and "experimental units" and provide the following two definitions.

> The units of statistical analysis are the data (the actual numbers) that we consider to be the outcomes of independent replications of our experiment. If you will, the units of statistical analysis are the numbers that we count when we count up degrees of freedom "within" or "for replication."
>
> The experimental units are the smallest division of the collection of experimental subjects that have been randomly assigned to the different conditions in the experiment and that have responded independently of each other for the duration of the experiment.

From these two definitions, it is apparent that if the assumption of statistical independence between students within a class is not appropriate then the unit of statistical analysis and the experimental units are clearly different.

Most teachers have examples of classes with distinct collective personalities and would freely agree with the Gestalt idea that a class as a whole is more than the sum of its students. All educators are aware of situations where one or two disruptive students cause the teacher to spend a considerable amount of class time managing these students in lieu of instruction. Or for a positive example, a class might have a very cooperative and supportive group of students and as a result accomplished far more than expected. In the above two cases, it is obvious that an individual student does not respond independently and consequently, the a priori assumption that each student responds independently is not reasonable.

Not only can student responses be dependent within classrooms, generally students are not randomly assigned to classes in the first place. The assignment of students to classes is a purposeful task performed by teachers and administrators and is not a haphazard or random activity. This deliberate assignment of students to classrooms fails to satisfy the definition presented earlier for students to be

considered the experimental units. In addition, Lumsdaine (1963) warns that even if students were assigned randomly to classes and treatments, important sources of error variation may not be accurately reflected in error estimates when the individual is used as the experimental unit. In a classical two-stage sampling design students are randomly assigned to classes and then classes are randomly assigned to treatments. Thus, at the outset of an experiment under these conditions the intraclass correlation is equal to zero. But during the duration of the treatment the teacher and students have the opportunity to interact, and this has the effect of inducing the intraclass correlation to be greater than zero.

A number of researchers have examined the unit of analysis question and its methodological implications. Lindquist (1953) advocated using the analysis of variance technique on the group means and has stimulated much subsequent research on the appropriate unit of analysis and associated statistical questions. Barcikowski (1981) extended Lindquist's work and investigated statistical power when group means were used as the unit of analysis with the analysis of variance method. The robustness of the t-test to violations of the assumptions regarding the unit of analysis were investigated by Blair, Higgins, Topping and Mortimer (1983). They found that even small amounts of between-class variation caused inflation in the type I error rate of the t-test when an analysis was incorrectly carried out at the student level. Further work on the specification of the proper statistical model that identifies the random factor(s) explicitly in the linear model was examined by Peckman, Glass and Hopkins (1969), and Hopkins (1982). This literature has emphasized the fact that the effects of intact groups cannot be ignored a priori. But, is the unit of analysis the critical question for the educational researcher?

Education as it takes place in our schools is distinguished by the fact that learning, an essentially psychological activity, is cultivated in a social environment. It is not sufficient for the researcher to choose between the student and the class as the appropriate unit of analysis. The principal consideration is the interplay between the psychological and social elements in our educational system. Teachers are faced with the difficult task of mediating and balancing individual student concerns with those of the class. Determining the optimal point between individual and group instruction that produces the maximum educational benefits is an important challenge. This interplay between the individual and the group must also be reflected in educational research methodology.

The differential responses of individuals to treatments is what Cronbach (1957) called "Aptitude X Treatment Interaction" (ATI). This differential effect a treatment has on individuals with different psychological characteristics depicts a frequent situation found in educational research. Educational research methods must be sensitive to the covariation between the psychological and social components of education and learning. Good estimates for the variance-covariance terms of the model are required to recognize important ATI effects.

The intraclass correlation coefficient is another measure that reveals a type of relationship between individual and group information. The intraclass correlation is constructed from a partitioning of the total random or unexplained variance into two components, one for the within-group variation and the other for the between-group variation. The intraclass correlation coefficient is then defined by the ratio of the between-group variance relative to the total variance, and as a consequence, its value ranges from zero to one. For the special case when the intraclass correlation

is considered to be known, Blair and Higgins (1986) recommend a weighted least squares solution to the unit of analysis problem. The conclusions reached using this analysis strategy are somewhere between those obtained using the individual as the unit of analysis and the group mean depending on the specific value of the intraclass correlation coefficient. Except for the degrees of freedom used to find the tabled critical value, an intraclass correlation approaching zero would produce results comparable to those using individuals as the unit of analysis and an intraclass correlation approaching unity would produce results comparable to using the group as the unit of analysis. Blair and Higgins (1986) mention that the intraclass correlations found in classroom level educational research have a rather narrow range. However, the intraclass correlation is generally unknown in practice and this implies that the required variance components must be estimated from the data. Consequently, their proposed solution is incomplete.

Not only are the variance components necessary for understanding the educational process under study, the covariance components also play an important part. In many problems in education, there are a number of variables that are of interest and these variables may be interdependent. Thus, a multivariate formulation is essential to model linear dependencies between variables and for joint probability statements about parameters that are linearly dependent. Raudenbush and Bryk (1988) provide an example where estimated parameter variance and covariance components enable a better understanding of school effects. In their example, they use the variance-covariance terms to construct the correlation coefficients between "excellence" (mean level of achievement) with "equity" as measured by the regression coefficients for minority status, social class, and academic background.

The Problem of an Unbalanced Design

Generally, the researcher has no control over the assignment of students to classes and there is nothing that prevents classes from varying in size. Not only does the number of students per class vary, there are other circumstances that contribute to an unbalanced design. If a research study is longitudinal in nature there is the problem of attrition. Students may be absent during the days designated for treatment or testing. In other cases, students and their families may move while a research project is in progress.

Statistical techniques are available for the balanced design, for example when there is the same number of students per class (Kirk, 1968; Winer, 1971; and Searle, 1971). When intact groups like classes are fundamental to educational research, the probability of obtaining a balanced design is extremely small. The application of many ANOVA type estimation procedures to unbalanced designs do not produce consistent results and as a consequence, results may vary dependent on the statistical procedure used or the particular manner it is applied (Searle, 1971).

There have been some recent advances in the application of statistical procedures to unbalanced designs. A number of researchers have successfully applied a variety of numerical maximum likelihood procedures to unbalanced hierarchical data found in educational research. The Fisher scoring method (Deleeuw and Kreft, 1986; and Longford, 1987), the iterative generalized least squares method (Goldstein, 1987) and the EM algorithm (Mason, Wong and Entwistle, 1984; Bryk, Raudenbush, Seltzer and Congdon, 1986) all produce maximum likelihood estimates with their desirable large sample properties. In particular, maximum likelihood estimators are asymptotically normal. Hence, for

large scale studies, maximum likelihood estimation provides a solution to the problems of unbalanced designs. But for small sample situations, the appropriateness of the large sample properties of maximum likelihood estimation is unclear.

Small Sample Problems

Along with the problems associated with intact groups, there are typically constraints placed on the resources available for educational research. Experimental research on education is expensive and compounds the problems associated with the hierarchical structure of educational data. For example, sampling considerations are more problematic in a hierarchical structure. Each level up the hierarchy represents a substantially smaller population. It is much easier to obtain the participation of twenty students than the same number of classes or even more difficult, to enlist the participation of twenty schools. This is especially true if the study requires the implementation of treatments. The recruitment and training of cooperating teachers, and the careful monitoring of the treatment delivery are very labor intensive activities.

Also, most educational treatments are implemented over a period of time and the data collection phase requires considerable resources over the course of the study. Multiple observations, testing and test scoring are all standard tasks in educational research. The sum of these necessary research activities makes it often impractical to have large-scale individual experimental research studies on education. It is more likely that smaller scale projects will continue to be the norm.

Small sample problems create additional demands for statistical procedures

not found with large samples. For example, the Central Limit Theorem states that the sampling distribution of the mean is approximately normal regardless of the population distribution. The only restrictions of this theorem is that the sample must be sufficiently large and the population variance must be finite. To evaluate the sampling distribution of a mean from a small sample, however, it is necessary to make some assumptions about the population distribution. To make inferences about the distribution of the sample mean based on a small sample, Student's t-test requires that the sample be drawn from a population with a normal distribution. Intuitively, it makes sense that the less information provided by the sample or data the more a researcher has to rely on prior knowledge and the validity of the assumptions.

A Review of Estimation Procedures

The hierarchical structure of educational data can be represented as a mixed-model, a subclass of linear models. A mixed-model conceptualization is appropriate because a combination of both fixed and random factors are included in the model. Fixed factors are factors in which all treatment levels or categories of interest are included in the experiment. Inferences about a fixed factor pertain only to the specific levels of the factor that have been included in the experiment. In research on education a fixed factor might be method of instruction, type of materials, or media used. Random factors have their associated treatment levels randomly sampled from the population of interest. Inferences about random factors are usually directed toward their population variances, and this information is often used to construct benchmarks to evaluate the relative size of fixed effects included in the

model. Students, classrooms, and schools are typical examples of random factors in educational research, and estimation of the variance and covariance components of these random factors is an important function in research on education.

In the context of mixed models, statistical estimation is difficult for unbalanced hierarchical designs where the interaction effect of measured student characteristics by treatments is an important focus of the investigation. This difficulty is compounded in small sample situations. Many estimation procedures that have analytical solutions for balanced designs fail for the unbalanced case and numerical approximation techniques are required. Some of the estimation procedures that have been proposed as solutions to the estimation problems presented by the combination of a hierarchical data structure, the interaction of measured student level characteristics by class level treatments, an unbalanced design, and a small sample size are reviewed next. These estimation procedures are divided into two groups, the analysis of variance (ANOVA) type procedures and the maximum likelihood procedures. I shall discuss the advantages and identify the drawbacks with the estimation procedures within these two groups.

The ANOVA Procedures

The problem of estimation for unbalanced mixed-model designs has a long history. R. A. Fisher is generally credited with systematically developing factorial designs along with the analysis of variance method. This standard ANOVA procedure provides the starting point for critically examining some of the alternative procedures proposed in this framework to handle the unbalanced case.

From his field work in animal science and the prevalence of unbalanced

experimental designs in this domain of research, Henderson (1953) proposed three methods for variance and covariance component estimation for this situation. The first method of estimation, referred to as Method 1, was essentially the conventional ANOVA method and is a starting point for comparisons. With this traditional ANOVA method, point estimates for the variance and covariance components in some situations can be calculated by equating the mean squares to their expected values. This set of equations can be solved for the unknown variance and covariance components in the balanced design case. In the unbalanced case, obtaining this solution is complicated by the increased complexity of the coefficients of each variance and covariance component. More importantly for unbalanced designs, Searle (1971, p. 41) states, "This is generally true; in mixed models, expected values of the S's ("analogous" sum of squares) contain functions of the fixed effects that cannot be eliminated by considering linear combinations of the S's." In this context, the S's are computed with unbalanced data in an analogous manner to the sum on squares computed with balanced data, but they do not necessarily have the same properties. Searle (1971, p. 36) provides an example where an S term is a quadratic form that is not positive definite. In effect this means that in many cases the variance and covariance components cannot be extracted using the ANOVA method.

Method 2 proposed by Henderson (1953) involves computing the least squares estimates for the fixed effects, correcting the data by subtracting the fixed effects, and then using Method 1 on the corrected data. Searle (1971) criticizes this method when used for mixed-model estimation on the grounds that it is not uniquely defined and it does not apply whenever the model includes interactions

between the fixed effects and the random effects. These two concerns expressed by Searle illustrate the problems with this method for mixed-models. Therefore this strategy is inappropriate for investigating the interaction between student level and class level variables.

The third method proposed by Henderson (1953), the fitting of constants method, also has some shortcomings. It uses reductions in the sums of squares due to fitting different submodels. These reduction terms are then set equal to their expected value and solved for the unknown variance and covariance components. This procedure does not produce a unique solution because the order in which the reduction terms are fitted may lead to different values for the associated mean squares. In most situations, there is no guiding principle for ordering the calculations of the reduction terms using this fitting of constants method, making this method unsatisfactory for the unbalanced mixed-model situation.

There is nothing inherent in these three estimation methods in the mixed-model case that prevents negative estimates of variance components in either the balanced or unbalanced cases (Searle, 1971). If a given application of one of these three methods does produce a negative estimate for a variance component, a researcher could set the estimate to zero. But how much confidence can be placed in an estimate that is outside the parameter space? In the balanced case, the estimates of variance components are unbiased for the three methods (Searle, 1971) and also Method 3 produces unbiased estimates in the unbalanced case. Since these unbiased variance estimates may be negative, the criteria for unbiased variance estimation may not be of primary importance.

Furthermore, Henderson (1953) states that in unbalanced design situations the

sampling variances of the estimates obtained by the three estimation methods are not known. As a result, hypotheses testing of the parameters in the model can not be evaluated with the same confidence as in a balanced design. This uncertainty about the sampling variances of the estimates also applies to statements concerning interval estimates.

The introduction of covariates, such as student abilities, aptitudes, and interest, create additional problems. The ANOVA procedure has been extended to include covariates. The analysis of covariance, ANCOVA, requires the assumption that the within-class regression coefficients are homogeneous (Winer, 1971). This assumption may not be tenable in many research situations. For example, it is quite possible that the effect of a covariate measuring a student background characteristic may vary across schools. In this case, the assumptions underlying the traditional ANCOVA procedure would be violated. When the assumption of homogeneous regression coefficients is not appropriate, the obvious question is why are the effects different. An alternative approach proposed by Cronbach and Webb (1975) was to estimate separate regressions within each group, but they concluded that when the sample size for each group was small the resulting estimates were not stable. Thus, the ANCOVA and the estimation of separate regressions methods may not provide a satisfactory solution for the heterogeneity of regressions problem. In addition, the ANCOVA procedure does not provide a general solution for estimating the covariance of a random student level effect with a random class level effect. A general solution must estimate this important covariance when the variables are measured on a discrete or continuous scale and is applicable to both balanced and unbalanced design situations.

Maximum Likelihood Procedures

The problems encountered with the ANOVA and ANCOVA estimation methods have led to the investigation of other estimation procedures. The maximum likelihood procedure of estimation, with the availability and computational speed of computers, has received renewed interest. This approach to estimation has many desirable features. One feature of this method is that it constrains the variance components to be nonnegative. Maximum likelihood estimators are functions of every sufficient statistic, consistent, asymptotically normal, and efficient (Harville, 1977). These properties of maximum likelihood estimators provide a large sample solution to the problems resulting from an unbalanced design. Generally, the distribution of maximum likelihood estimators in small sample situations are unknown.

Implementation of the maximum likelihood method is not always easily accomplished. In the unbalanced mixed-model situation with the variance-covariance components unknown, there is no general closed form or explicit analytic solution to the maximum likelihood equations (Searle, 1971). A numerical procedure must be used in this situation. One of the first numerical methods used for maximum likelihood estimation was the Newton-Raphson method. This method has the desirable property of quadratic convergence. The disadvantage of the Newton-Raphson method is that there is no guarantee that it will converge to the solutions of the maximum likelihood equations. It might converge to a local rather than the global solution (Harville, 1977). This method may depend on good initial approximations. In some cases, if the initial approximations are not sufficiently close to the actual roots of the likelihood equations, the Newton-Raphson's method

may diverge (Harville, 1977; Burden, Faires and Reynolds, 1981).

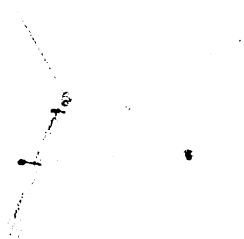
There are a number of alternative computational approaches to the Newton-Raphson method that produce maximum likelihood estimates of the parameters of a hierarchical linear model. The Fisher scoring method (DeLeeuw and Kreft, 1986; Longford, 1987), iterative generalized least squares (Goldstein, 1987), and the EM algorithm (Mason, Wong and Entwistle, 1984; Bryk, Raudenbush, Congdon and Seltzer, 1986) are examples of different computational approaches currently available to researchers that produce maximum likelihood estimates. The EM algorithm developed by Dempster, Laird and Rubin (1977), and implemented with the HLM Computer Program (Bryk et al., 1986) was selected for use in this study. This estimation procedure has been applied to a variety of research problems in education with success. For examples of this algorithm used in educational research, see Dempster, Rubin and Tsutakawa (1981), Laird and Ware (1982), Strenio, Weisberg and Bryk (1983), and Raudenbush and Bryk (1985, 1986). This algorithm gets its name from the two basic iterative steps: the expectation step and the maximization step. Wu (1983) has shown that the EM algorithm will converge when the complete data are from a curved exponential family, and the expected log of the likelihood function is unimodal and satisfies a mild differentiability condition. Generally, the EM algorithm has a slower rate of convergence than the Newton-Raphson method. However, the EM algorithm will converge when given well-defined initial estimates of the parameters along with reasonable distribution assumptions satisfying the conditions outlined above. The availability of this software package and its superior convergence property were the primary reasons for selecting this method.

An advantage of maximum likelihood estimation is that it may be applied to the hierarchical linear model to produce empirical Bayes estimates. The hierarchical model may be easily formulated to produce empirical Bayes estimates which are equivalent to restricted maximum likelihood estimates (Laird and Ware, 1982). Harville (1977) points out that one criticism of the standard maximum likelihood approach to estimation is that it does not take into account the loss in degrees of freedom that results from estimating the fixed effects of the model. He points out that inadequacies of this traditional maximum likelihood (ML) approach are overcome by restricted maximum likelihood (REML) estimation because this approach does take into account the loss of degrees of freedom due to estimating fixed effects (Harville, 1977). Dempster, Rubin and Tsutakawa (1981) differentiate the ML and REML approaches by denoting the two likelihood functions by MLF and MLR, where F stands for fixed and R for random.

The empirical Bayes approach capitalizes on the strengths of the data. For example, in a two-level hierarchical linear model with student and class levels, regression coefficients representing the regression of an outcome variable on one or more student level independent variables may be expressed as a precision-weighted combination of the contributions for the entire data set and the individual class. This method uses the strengths of the data for estimation, because the more precisely a component of the model is measured the more weight it will receive in constructing an estimate. Thus, a regression equation for a particular classroom containing a small number of students may borrow information across classes to obtain an improved estimate. This estimation approach provides a method to resolve the problem of estimating individual regression equations from nested units

with small samples. The details of the precision-weighting scheme are presented in Chapter III.

The empirical Bayes approach, in the context of the hierarchical linear model, is available to educational researchers with the HLM computer program (Bryk, et al., 1986). This method of estimation is most suited to large sample problems. The distribution of the estimates resulting from the EM algorithm applied to a hierarchical model are generally not known in small sample situations. For example, in a hierarchical model with students nested in classes, a sufficiently large sample of classes is required for trustworthy results. In addition, the standard errors for any fixed effects at the class level fail to take into account the uncertainty associated with the estimation of the variance-covariance components, which is an important consideration in small sample problems. A basic strategy of the maximum likelihood methods is to calculate an estimate of the variance-covariance components and use this point estimate as if it were known. For small sample problems, variance-covariance estimates may be quite unstable. The small sample dispersion of maximum likelihood estimates that rely on variance-covariance estimates do not reflect the consequences of the variation that may result from other plausible variance-covariance component values. The empirical Bayes approach relies on point estimates of the variance-covariance components to construct weights and there has been some evidence (Bassiri, 1988) that small sample empirical Bayes estimates have standard errors that are two small and do not reflect the uncertainty associated with the estimation of the required variance-covariance components.



A Proposed Solution

An alternative methodology that may be applied to this analysis problem is a Bayesian approach. The Bayesian method is not new and was initially brought into focus by Reverend Thomas Bayes (1763). The application of true Bayesian methods to hierarchical models has been impeded because the mathematics involved has been intractable. Lindley and Smith (1972) indicate that most reasonable distributions employed in a hierarchical model lead to integrals which cannot all be expressed in closed form. The Bayesian approach is faced with the problem that the required equations cannot always be expressed in closed form. The EM algorithm supplied an iterative solution to this problem for the maximum likelihood method. Tanner and Wong (1987) have developed a new data augmentation algorithm for the implementation of the Bayesian method. This method provides an innovative means for the computation of posterior distributions.

The key advantage of the Bayesian approach is that it supplies the joint posterior distribution of the parameters of the model. The joint posterior distribution provides more information to the researcher than joint modal estimates for the parameters produced by the empirical Bayes approach. For example, it is possible for a researcher to calculate the mode, mean, median, and percentile points for any parameter of interest from the joint posterior distribution. In addition, small sample parameter estimates may be unstable and cause problems when they are considered known in empirical Bayes or other maximum likelihood methods. The Bayesian approach uses distributions rather than point estimates and the joint posterior distribution resulting from this approach reflects the uncertainty of our knowledge about the parameters in the model. For example, variance-covariance

parameters are assumed to have a distribution and their contribution to the joint posterior distribution is determined by including the probability of all possible values across their distribution. Hence, if different plausible values of the variance-covariance parameters result in different estimates from the empirical Bayes approach, this variability is reflected in the Bayesian approach. The empirical Bayes approach will have a tendency to underestimate this variability. Since many research projects in education are conducted with small to moderate sample sizes, especially when the data has a hierarchical structure, the Bayesian approach may provide a solution to small sample size problems that have not been resolved by the empirical Bayes approach.

In this study, the Bayesian approach implemented with the data augmentation algorithm is applied to a common experimental design found in research on education. This design includes an independent variable measured at the student level. This independent variable could be a measured student characteristic such as motivation, aptitude, or interest. In addition, a simple experimental design representing an experimental group and a control group was defined at the classroom level. The specific details of this model are presented in the next chapter.

Hierarchical data illustrating representative samples found in research on education at the class level are simulated with known properties. A number of distinct data sets are generated with different combinations of between-class effect parameters. The number of students per class is drawn randomly from a realistic distribution of class sizes that might be found at the elementary school level. The remainder of the parameters that are necessary to define the simulated data are set

to values typically obtained from research on education.

Since the data augmentation algorithm is new, there are questions concerning the most effective implementation of the algorithm and the amount of computational resources it requires. The data augmentation method requires the generation of latent data that is added to the observed data, as its name implies. The amount of data generated is a variable of the algorithm and a technique for controlling this aspect must be determined before it can be implemented. A strategy must be developed to obtain an optimal balance between the amount of data generated and the computational resources required. Furthermore, this algorithm is an iterative procedure and stopping rules must be determined prior to programming it for the computer. The practicality of this algorithm in terms of the necessary computational resources needs to be evaluated due to the combination of data generation and iteration. This is an important consideration because in substantive educational research situations the data are analyzed a number of times using different models. Any data analysis procedure should be efficient enough to make this flexibility practical.

The investigation of the performance of the implementation of the Bayesian statistical procedure is another important consideration. A practitioner must know the properties of a statistical procedure. In addition, researchers are interested in knowing which estimation method produces the best results relative to a set of possible methods currently available. To make a meaningful comparison the empirical Bayes approach using the EM algorithm was selected as an alternative estimation procedure. The performance of these methods under small sample conditions are investigated. A simulation study is unique because it provides

knowledge of the population parameters. A comparison of both methods can be made as to how accurately they recover the population parameters.

In a hypothesis testing context, type I error rates are compared against the nominal level for the relevant between-class level parameters. In addition, the type I error rates produced by the two statistical methods will be compared. A related concern is the statistical power of the test. The power of the two methods is estimated and compared.

CHAPTER II

NOTATION CONVENTIONS AND THE MODEL

This chapter provides a brief discussion of the matrix symbols and conventions employed in this study. Also, the specific hierarchical linear model used is first described in scalar terms and then extended to matrix notation. The basic assumptions for the empirical Bayes approach using the EM algorithm and the Bayesian approach using the data augmentation algorithm are presented.

The educational context features students nested within classes for the first level of the hierarchy, the within-class level. Let the dependent variable be a measure of student achievement in a subject area and the independent variable be a measured psychological attribute, such as an aptitude. The dependent variable is regressed on the independent variable resulting in two regression coefficients for each class, a slope and an intercept. At the between-class level, classes are randomly assigned to either the experimental treatment or the control group. At this level, the within-class parameters are considered random outcomes and are modeled as functions of the fixed effects due to treatments. In this example, it is the researcher's task to explain the variation among the outcomes of each class as a result of the between-class model's experimental treatment. This hierarchical linear model is translated into an explicit mathematical presentation.

A Guide to the Matrix Notation

Bold letters are used to represent matrices and vectors. Bold upper case letters are used to represent matrices and bold lower case letters are used to represent vectors. For example, **X** represents a matrix and **x** represents a vector. Furthermore, scalar values are represented by characters or numbers not in bold face type. The order of matrices and vectors is indicated by "(r x c)" to specify the number of rows and columns respectively. A matrix with r rows and c columns may be written as **X** (r x c) and displayed as

$$\mathbf{X} = \begin{vmatrix} X_{11} & X_{12} & \cdots & X_{1c} \\ X_{21} & X_{22} & \cdots & X_{2c} \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ \cdot & & & \cdot \\ X_{r1} & X_{r2} & \cdots & X_{rc} \end{vmatrix}.$$

An element of **X** in row i and column j is X_{ij}. In a similar manner, the (r x 1) column vector **y** may be illustrated as

$$\mathbf{y} = \begin{vmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ y_r \end{vmatrix}.$$

To help differentiate symbols, parameters are represented by Greek characters and Roman characters are used to represent observed values or known constants.

Moreover, the superscript "-1" is used to indicate the inverse of a matrix, an

apostrophe " ' " is used to indicate the transpose of a matrix. The identity matrix is denoted by **I**. "Var" is the variance operator, "Cov" is the covariance operator, "D" indicates a dispersion matrix, "Tr" is the trace operator, and "E" is the expectation operator. The symbol "$\propto$" indicates proportionality, and "~" can be read as, "is distributed." The following summarizes these conventions:

α	is a vector of parameters,
$\mathbf{W}$	is a matrix of fixed constants,
$\mathbf{A}^{-1}$	is the inverse of matrix **A**,
$\mathbf{A}'$	is the transpose of matrix **A**,
$\mathbf{I}$	is the identity matrix,
Var(x)	is the variance of x,
Cov(**x**, **y**)	is the covariance matrix of **x** and **y**,
D(**y**)	is the dispersion of the vector **y**,
Tr(Σ)	is the trace of the matrix Σ,
E(**y**)	is the expectation of the vector **y**,
$\mathbf{A} \propto c\mathbf{A}$	indicates the matrix **A** is proportional to c**A**, and
$y \sim N(\mu, \sigma^2)$	indicates that y is distributed normally, with mean μ and variance σ^2.

The Hierarchical Linear Model in Scalar Terms

At the first level, the within-class model specifies a model for the units of statistical analysis. For classroom level research in education, the first level involves students within a class. In this case, the within-class model includes two

parameters, an intercept and a slope. In other educational research situations this first level of the hierarchy might model classrooms nested within schools. Since two unique regression coefficients are determined for each class, the resulting set of regression parameters vary randomly across classes. The between-class model is developed to explain the random variation of these regression parameters.

More specifically, the within-class model, is presented in scalar terms as follows:

$$y_{ij} = \beta_{j0} + \beta_{j1}X_{ij1} + r_{ij}, \tag{2.01}$$

where

y_{ij} is the outcome response of the i^{th} student in the j^{th} class,

β_{j0} is the intercept (base) of the j^{th} class,

β_{j1} is the regression slope for the j^{th} class,

X_{ij1} is i^{th} student's predictor response, and

r_{ij} is a random error for the i^{th} student in the j^{th} class,

for

$i = 1, 2, \ldots, n_j$ with n_j students in the j^{th} class, and

$j = 1, 2, \ldots, k$ and k is equal to the number of classes.

At the second level, the between-class level, we specify a model for the random parameters of the first level. The random parameters, the intercepts and the slopes, are now considered to be outcomes and are explained as functions of between-class differences. These between-class differences could be explained with a design on the classes that includes experimental treatments, naturally occurring

classification factors, measured attributes of the classes, or some combination of these different types of independent variables. The between-class design selected for this study differentiates classes into experimental treatment and control groups.

The between-class model is presented in scalar terms as follows:

$$\beta_{j0} = \gamma_{00} + \gamma_{01}W_j + u_{j0}, \tag{2.02a}$$

$$\beta_{j1} = \gamma_{10} + \gamma_{11}W_j + u_{j1}, \tag{2.02b}$$

where

γ_{00} is the base of class intercepts,

γ_{01} is the main effect due to experimental treatment,

u_{j0} is a random error associated with the j^{th} class,

γ_{10} is the base of the within-class slopes,

γ_{11} is the interaction effect (ATI),

u_{j1} is a random error associated with the j^{th} class, and

W_j is equal to 1 for classes in the experimental group, and
is equal to -1 for the classes in the control group.

Equations (2.01), (2.02a) and (2.02b) may be combined to form a more traditional representation of the hierarchical model. This equivalent representation of the model may be written as

$$y_{ij} = \gamma_{00} + \gamma_{01}W_j + u_{0j} + \gamma_{10}x_{ij1} + \gamma_{11}W_jx_{ij1} + u_{1j}x_{ij1} + r_{ij}. \tag{2.03}$$

This traditional representation obscures the hierarchical structure. The terms in the

model are the same as described previously.

In many situations the unconditional model provides some valuable information. The unconditional model in this hierarchical context refers to the model without the inclusion of any between-class predictor variable(s), which in this case is the treatment indicator variable W_j. When both the unconditional and conditional models have been estimated the proportion of additional between-class variance accounted for by the inclusion of one or more predictors is available to the researcher to help assess the performance of the conditioning variable(s). The unconditional between-class model is presented in scalar terms below:

$$\beta_{j0} = \gamma_0 + u_{j0}, \tag{2.04a}$$

$$\beta_{j1} = \gamma_1 + u_{j1}, \tag{2.04b}$$

where

γ_0 is the base of the class intercepts,

u_{j0} is a random error associated with the j^{th} class,

γ_1 is the base of the within-class slopes, and

u_{j1} is a random error associated with the j^{th} class.

A Matrix Representation of the Hierarchical Linear Model

The next step is to extend the scalar representation of the model into matrix terms. In general, the within-class model may contain random effects, fixed effects, or a combination of these two types of effects. In this specific model, the within-class slope and intercept parameters are considered as random variables. The within-class model is given in matrix terms as follows:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{r}, \tag{2.05}$$

where

$$\mathbf{y} = \begin{vmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ . \\ . \\ . \\ \mathbf{y}_k \end{vmatrix} \quad (\Sigma n_j \times 1),$$

$$\mathbf{X} = \begin{vmatrix} \mathbf{X}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{X}_2 & \dots & \mathbf{0} \\ . & . & \dots & . \\ . & . & \dots & . \\ . & . & \dots & . \\ \mathbf{0} & . & \dots & \mathbf{X}_k \end{vmatrix} \quad (\Sigma n_j \times 2k),$$

with

$$\mathbf{X}_j = \begin{vmatrix} 1 & X_{1j1} \\ 1 & X_{2j1} \\ . & . \\ . & . \\ . & . \\ 1 & X_{n_j j1} \end{vmatrix} \quad (\Sigma n_j \times 2) \text{ and note that } X_{ij0} = 1,$$

$$\beta = \begin{vmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{vmatrix} \quad (2k \times 1),$$

$$\beta_j = \begin{vmatrix} \beta_{j0} \\ \beta_{j1} \end{vmatrix} \quad (2 \times 1),$$

$$\mathbf{r} = \begin{vmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ . \\ . \\ . \\ \mathbf{r}_k \end{vmatrix} \quad (\Sigma n_j \times 1),$$

with

$$\mathbf{r}_j = \begin{vmatrix} r_{1j} \\ r_{2j} \\ . \\ . \\ . \\ r_{n_j} \end{vmatrix} \quad (\Sigma n_j \times 1),$$

for

$j = 1, 2, \ldots, k$ with k equal to the number of classes, and

$i = 1, 2, \ldots, n_j$ for the students in the j^{th} class.

The between-class model, which comprises the second level in the hierarchy, is presented as follows:

$$\beta = \mathbf{W}\gamma + \mathbf{u}, \tag{2.06}$$

where β is defined as above and

$$\mathbf{W} = \begin{vmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ . \\ . \\ . \\ \mathbf{W}_k \end{vmatrix} \quad (2k \times 4),$$

and for a class in the experimental group

$$\mathbf{W}_j = \begin{vmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{vmatrix},$$

or for a class in the control group

$$\mathbf{W}_j = \begin{vmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{vmatrix},$$

$$\gamma = \begin{vmatrix} \gamma_{00} \\ \gamma_{01} \\ \gamma_{10} \\ \gamma_{11} \end{vmatrix} \quad (4 \times 1), \text{ and}$$

$$\mathbf{u} = \begin{vmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ . \\ . \\ . \\ \mathbf{u}_k \end{vmatrix} \quad (2k \times 1),$$

with

$$\mathbf{u}_j = \begin{vmatrix} u_{j0} \\ u_{j1} \end{vmatrix} \quad (2 \times 1).$$

Assumptions for the Two Statistical Procedures

For the empirical Bayes procedure, the following assumptions are made:

$$\mathbf{r} \sim N(\mathbf{0}, \Sigma), \text{ and } \mathbf{u} \sim N(\mathbf{0}, T). \tag{2.07}$$

In addition, we assume a noninformative prior distribution for the between-class parameter vector γ, or in other words, our knowledge about the actual parameter values of this distribution is almost null or very small. This noninformative prior is specified as

$$\gamma \sim N(\theta, \Gamma). \tag{2.08}$$

Furthermore, these simplifying assumptions are made:

$$\Sigma = \mathbf{I}\sigma^2, \tag{2.09}$$

and

$$\text{Cov}(\mathbf{r}, \mathbf{u}) = \mathbf{0}. \tag{2.10}$$

The assumption given in (2.10) implies that individual student errors are independent of class level errors and T is block diagonal with each T_j represented as

$$T_j = \begin{vmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{vmatrix}. \tag{2.11}$$

Furthermore, the T_j are assumed to be the same for all j, which is often referred to as homogeneity of variance-covariance matrices. Consequently, a pooled estimate of T is calculated for this common variance-covariance matrix.

The classes have been randomly assigned to either the experimental treatment or the control group. This is a common method used to control for existing differences at the class level prior to implementing treatments. Given that the classes were randomly assigned to the two experimental conditions, the model developed in (2.01), (2.02a) and (2.02b) is parsimonious. An alternative possibility that may occur when classes are not randomly assigned to treatments is that a linear relationship might exist prior to treatment between the class mean on the independent variable and the outcome measure. In this slightly more complicated situation, an extra term in the between-group model would be required to account for this effect.

The full Bayesian model for the data augmentation algorithm requires additional distribution assumptions for the parameters σ^2 and T. Conjugate prior

distributions are specified for these parameters. The inverse chi-square prior distribution for σ^2 is given by

$$\sigma^2 \sim \nu_0\sigma_0^2\chi^{-2}(\nu_0) \tag{2.12}$$

with the degrees of freedom parameter ν_0 and σ_0^2 indicates the prior value for σ^2. And finally, the prior distribution of T is an inverse Wishart distribution given by

$$T \sim W^{-1}(\Psi, \upsilon), \tag{2.13}$$

where Ψ is the parameter matrix of the inverse Wishart distribution and υ is the degrees of freedom parameter. Anderson (1984, p. 268) calls Ψ the "precision matrix." The degree of knowledge concerning the actual values of the parameters in (2.08), (2.12), and (2.13) and the contribution they have to the joint posterior distribution is assumed almost null or very small so that these prior distributions are noninformative. This means that before a sample is drawn the actual values of γ, σ^2 and T are vague. There are no specific values for γ, σ^2 or T within their respective parameter spaces that are considered more probable than any others before sampling. Any inference based on the posterior distribution, which is a combination of the data and the prior distributions, will be dominated by the information derived from the data.

CHAPTER III

APPLICATION OF THE EM AND DATA AUGMENTATION ALGORITHMS TO THE HIERARCHICAL LINEAR MODEL WITH NORMAL ERRORS

The formal derivations of the two algorithms are presented for the specific hierarchical linear model used in this simulation study. This model includes different design elements so that the application of these statistical procedures to other related designs is straightforward.

The Empirical Bayes Procedure Using the EM Algorithm

The empirical Bayes procedure presented in this section draws heavily on prior research. The following works have been especially important: Lindley and Smith (1972) and Smith (1973) for their work putting the hierarchical linear model in a Bayesian framework; Dempster, Laird and Rubin (1977) for developing the EM algorithm; Strenio, Weisberg and Bryk (1983) and Raudenbush (1984) for the application of this approach in an educational context; and the development of the HLM computer program by Bryk, Raudenbush, Seltzer and Congdon (1986). Furthermore, the HLM computer program was used to produce the empirical Bayes estimates for this study.

The two level hierarchical linear model and its assumptions for the empirical Bayes approach are restated next:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{r}, \qquad \mathbf{r} \sim N(\mathbf{0}, \Sigma),$$

$$\beta = \mathbf{W}\gamma + \mathbf{u}, \qquad \mathbf{u} \sim N(\mathbf{0}, T),$$

$$\gamma \sim N(\theta, \Gamma),$$

and

$$\Sigma = I\sigma^2 \quad \text{and} \quad \text{Cov}(\mathbf{r}, \mathbf{u}) = \mathbf{0}.$$

Initially, Σ and T are assumed known. The goal of the empirical Bayes approach is to estimate the joint posterior modes of β and γ given $\mathbf{y}$, Σ and T. This joint posterior distribution is given by

$$p(\mathbf{y}, \beta, \gamma \mid \Sigma, T) \propto f(\mathbf{y} \mid \beta, \Sigma)\, g(\beta \mid \gamma, T)\, h(\gamma). \tag{3.01}$$

The term empirical Bayes is used to describe this procedure because the β parameters at the first level use the second level of the model as a prior, which is a Bayesian idea. The second level does not use an informative prior distribution for γ. Consequently, the second level parameters are estimated empirically from the data. No prior distributions are specified for Σ and T. From a Bayesian perspective, the result of this statistical procedure is the calculation of the joint posterior modes of the two parameter vectors β and γ from (3.01).

To determine the empirical Bayes estimator for β and γ, the partial derivatives of (3.01) must be taken with respect to β and γ. The prior distribution $h(\gamma)$ is considered a very small constant and it can be ignored while the empirical Bayes estimators are calculated. The resulting equations are then set to zero and

solved to provide the empirical Bayes estimators. To begin with,

$$p(\beta, \gamma \mid y, \Sigma, T) \propto \exp\{-1/2(y - X\beta)'\Sigma^{-1}(y - X\beta) - 1/2(\beta - W\gamma)'T^{-1}(\beta - W\gamma)\}. \quad (3.02)$$

After taking the natural logarithm of (3.02), expanding and collecting the terms in the exponent, and eliminating all terms not containing β or γ, the following expression results from these operations:

$$\text{Ln } p(\beta, \gamma \mid y, \Sigma, T) \propto \beta'X'\Sigma^{-1}X\beta - 2\beta'X'\Sigma^{-1}y + \beta'T^{-1}\beta - 2\gamma'W'T^{-1}\beta + \gamma'W'T^{-1}W\gamma. \quad (3.03)$$

Next, the partial derivative of the logarithm of expression (3.03) is taken with respect to β:

$$\partial \text{Ln } p(\beta, \gamma \mid y, \Sigma, T) / \partial\beta = 2X'\Sigma^{-1}X\beta - 2X'\Sigma^{-1}y + 2T^{-1}\beta - 2T^{-1}W\gamma. \quad (3.04)$$

Setting (3.04) equal to zero and solving for β, assuming all matrices that are inverted are non-singular, results in

$$\beta = (X'\Sigma^{-1}X + T^{-1})^{-1} (X'\Sigma^{-1}y + T^{-1}W\gamma). \quad (3.05)$$

To help simplify (3.05), the following terms are defined:

$$\mathbf{L} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{T}^{-1}), \tag{3.06}$$

$$\mathbf{V} = (\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{X})^{-1}, \tag{3.07}$$

$$\boldsymbol{\Lambda} = \mathbf{L}^{-1}\mathbf{V}^{-1}, \tag{3.08}$$

and

$$\mathbf{I} - \boldsymbol{\Lambda} = \mathbf{L}^{-1}\mathbf{T}^{-1}. \tag{3.09}$$

With the above definitions, equation (3.05) can now be simplified as follows:

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{L}^{-1}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}) + \mathbf{L}^{-1}\mathbf{T}^{-1}\mathbf{W}\boldsymbol{\gamma} \\ &= \mathbf{L}^{-1}\mathbf{V}^{-1}\mathbf{V}(\mathbf{X}'\boldsymbol{\Sigma}^{-1}\mathbf{y}) + (\mathbf{I} - \boldsymbol{\Lambda})\mathbf{W}\boldsymbol{\gamma} \\ &= \boldsymbol{\Lambda}\boldsymbol{\beta}^{0} + (\mathbf{I} - \boldsymbol{\Lambda})\mathbf{W}\boldsymbol{\gamma}. \end{aligned} \tag{3.10}$$

Note that $\boldsymbol{\beta}^0$ in equation (3.10) is an ordinary least squares estimator of the within-class parameter vector $\boldsymbol{\beta}$ when $\boldsymbol{\Sigma} = \mathbf{I}\sigma^2$. An estimator for $\boldsymbol{\gamma}$ is still required to determine the empirical Bayes joint modal estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ given $\mathbf{y}$, $\boldsymbol{\Sigma}$ and T.

The modal estimator for the $\boldsymbol{\gamma}$ parameter vector is determined similarly to the estimator for $\boldsymbol{\beta}$. First, the partial derivative of the natural logarithm of (3.01) is taken with respect to $\boldsymbol{\gamma}$ and then the result is set equal to zero. These operations result in

$$\partial \text{Ln } p(\boldsymbol{\beta}, \boldsymbol{\gamma} \mid \mathbf{y}, \boldsymbol{\Sigma}, \mathbf{T}) \,/\, \partial\boldsymbol{\gamma} = 2\mathbf{W}'\mathbf{T}^{-1}\mathbf{W}\boldsymbol{\gamma} - 2\mathbf{W}'\mathbf{T}^{-1}\boldsymbol{\beta}, \tag{3.11}$$

and

$$\boldsymbol{\gamma} = (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{T}^{-1}\boldsymbol{\beta}. \tag{3.12}$$

The results of (3.10) can be substituted into (3.12) as follows:

$$\gamma = (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{T}^{-1}[\Lambda\beta^{0} + (\mathbf{I} - \Lambda)\mathbf{W}\gamma]. \tag{3.13}$$

Multiplying out (3.13), rearranging some terms, and then multiplying through by $(\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})$ results in

$$[\mathbf{I} - (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}(\mathbf{W}'\mathbf{T}^{-1}\mathbf{W}) + (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{T}^{-1}\Lambda\mathbf{W}]\gamma \tag{3.14}$$
$$= (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{T}^{-1}\Lambda\beta^{0},$$

and

$$\mathbf{W}'\mathbf{T}^{-1}\Lambda\mathbf{W}\gamma = \mathbf{W}'\mathbf{T}^{-1}\Lambda\beta^{0}.$$

Two helpful identities (identities 3 and 4 from Smith, 1973) can be used to rewrite equation (3.14) in a more informative representation. Let

$$\begin{aligned}\Delta^{-1} &= (\mathbf{T} + \mathbf{V})^{-1} \qquad (3.15)\\ &= \mathbf{T}^{-1} - \mathbf{T}^{-1}(\mathbf{T}^{-1} + \mathbf{V}^{-1})^{-1}\mathbf{T}^{-1}\\ &= \mathbf{T}^{-1} - \mathbf{T}^{-1}[\mathbf{I} - (\mathbf{T}^{-1} + \mathbf{V}^{-1})^{-1}\mathbf{V}^{-1}]\\ &= \mathbf{T}^{-1} - \mathbf{T}^{-1} + \mathbf{T}^{-1}[(\mathbf{T}^{-1} + \mathbf{V}^{-1})^{-1}\mathbf{V}^{-1}]\\ &= \mathbf{T}^{-1}[(\mathbf{T}^{-1} + \mathbf{V}^{-1})^{-1}\mathbf{V}^{-1}].\end{aligned}$$

From (3.08)

$$\Lambda = \mathbf{L}^{-1}\mathbf{V}^{-1} = (\mathbf{T}^{-1} + \mathbf{V}^{-1})^{-1}\mathbf{V}^{-1},$$

and hence,

$$\Delta^{-1} = \mathbf{T}^{-1}\Lambda.$$

Substituting Δ^{-1} for $T^{-1}\Lambda$ in (3.14) and solving for γ results in the empirical Bayes estimator

$$\gamma^* = (\mathbf{W}'\Delta^{-1}\mathbf{W})^{-1}\mathbf{W}'\Delta^{-1}\beta^0, \tag{3.16}$$

and it follows from (3.10) that

$$\beta^* = \Lambda\beta^0 + (\mathbf{I} - \Lambda)\mathbf{W}\gamma^*. \tag{3.17}$$

The empirical Bayes estimates of γ^* and β^* in (3.16) and (3.17) are defined as the values of γ and β in their respective parameter spaces that jointly maximize expression (3.02).

Smith (1973) presents the posterior dispersions of γ^* and β^* conditional on $\mathbf{y}$, Σ and T. These dispersion matrices are

$$D(\gamma^*) = (\mathbf{W}'\Delta^{-1}\mathbf{W})^{-1}, \tag{3.18}$$

and

$$D(\beta^*) = \Lambda\mathbf{V} + (\mathbf{I} - \Lambda)D(\mathbf{W}\gamma^*)(\mathbf{I} - \Lambda), \tag{3.19}$$

where

$$D(\mathbf{W}\gamma^*) = \mathbf{W}(\mathbf{W}'\Delta^{-1}\mathbf{W})^{-1}\mathbf{W}'.$$

When the variance-covariance components are unknown, there is no general analytical solution to this estimation problem. The EM algorithm provides one iterative, numerical solution.

The EM Algorithm and Covariance Estimation

The empirical Bayes estimator for β and γ presented in the preceding section assume that Σ and T are known, a situation rarely if ever found in practice. These variance-covariance matrices must be estimated from the data. The empirical Bayes approach to this problem is to substitute maximum likelihood estimates of Σ and T into the expression (3.02). The result of this approach is to obtain the joint modal posterior distribution of β and γ given **y** and the variance-covariance matrices Σ and T are equal to their maximum likelihood estimates. To begin our development, let us consider the two random vectors **r** and **u** had been observed. In this case, maximum likelihood estimates for Σ and T could be calculated easily.

The structures of Σ and T implied by the assumptions reduce the size of the matrices that are required to compute the estimates of these two variance-covariance matrices. Since it was assumed that $\text{Cov}(\mathbf{r},\mathbf{u}) = \mathbf{0}$ and $\Sigma = \mathbf{I}\sigma^2$, a pooled within-class estimate can be obtained for the variance-covariance matrix Σ. Also, it has been assumed that the β_j's are independently and identically distributed and the sufficient statistic for T can be pooled across classes by summing the submatrices on the diagonal. Let $\mathbf{r}'\mathbf{r}$ and $\mathbf{u}_j\mathbf{u}_j'$ be the sufficient statistics for the two variance-covariance matrices. It follows that the maximum likelihood estimates for σ^2 and T are

$$\sigma^2 = \mathbf{r}'\mathbf{r} \,/\, (\Sigma n_j), \tag{3.20}$$

and

$$T = (1/k)\ \Sigma \mathbf{u}_j\mathbf{u}_j'. \tag{3.21}$$

The central idea behind the EM algorithm is to estimate the sufficient statistics that are required to calculate the maximum likelihood estimates for σ^2 and T as presented in equations (3.20) and (3.21). The EM algorithm has two basic steps, the expectation and the maximization steps. The basic outline of the EM algorithm is presented below.

1. The expectation step (E-Step) requires initial estimates for β^* and γ^*. These estimates are then considered as if they were known parameters and used to calculate the conditional expectations of the sufficient statistics given **y**, Σ and T necessary to construct the sum of squares associated with Σ and T.
2. The maximization step (M-Step) uses the new estimated sum of squares derived in step 1 for Σ and T to calculate updated maximum likelihood estimates for these two variance-covariance matrices.
3. The updated estimates for the variance-covariance matrices generated in the maximization step are now recycled back to the E-Step to recalculate new values for β^* and γ^*. The newest values for β^* and γ^* are used to generate another set of expected sum of squares for the variance-covariance matrices, which are then passed on to the maximization step. This process continues to iterate between the expectation and maximization steps until a stopping criterion is reached. One criterion is to stop the iterative process when the change in a function of the joint posterior distribution between successive iterations is less than some specified tolerance.

The strategy employed by the EM algorithm is to compute the conditional expectations of $\mathbf{r'r}$ and $\sum \mathbf{u}_j\mathbf{u}_j'$ given $\mathbf{y}$, Σ and T. These estimated sufficient statistics are then passed to the maximization step to estimate Σ and T. It has been assumed that $\Sigma = \mathbf{I}\sigma^2$ and $\text{Cov}(\mathbf{r}, \mathbf{u}) = \mathbf{0}$. Let the superscript "p" indicate the previous iteration's posterior parameter estimates. The previous iteration's posterior estimates for the variance-covariance matrices Σ and T are considered known for the purpose of calculating the conditional expectations. The following details for determining the required conditional expectations are from Dempster, et al. (1981). First the conditional expectation for the estimated sufficient statistic for $\Sigma = \mathbf{I}\sigma^2$ given $\mathbf{y}$, Σ and T is

$$
\begin{aligned}
E(\mathbf{r'r}) &= E(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \qquad (3.22)\\
&= E(\mathbf{y} - \mathbf{X}\beta^p + \mathbf{X}\beta^p - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta^p + \mathbf{X}\beta^p - \mathbf{X}\beta)\\
&= (\mathbf{y} - \mathbf{X}\beta^p)'(\mathbf{y} - \mathbf{X}\beta^p) + E(\beta - \beta^p)'\mathbf{X'X}(\beta - \beta^p)\\
&= (\mathbf{y} - \mathbf{X}\beta^p)'(\mathbf{y} - \mathbf{X}\beta^p) + \text{Tr}[\mathbf{X'X}\ \text{Var}(\beta - \beta^p)]\\
&= (\mathbf{y} - \mathbf{X}\beta^p)'(\mathbf{y} - \mathbf{X}\beta^p) + \text{Tr}[\mathbf{X'X}\ \text{Var}(\beta)],
\end{aligned}
$$

with

$$
\begin{aligned}
\text{Var}(\beta) &= \text{Var}(\mathbf{W}\gamma) + \text{Var}(\mathbf{u}) + \text{Cov}(\mathbf{W}\gamma, \mathbf{u}) + \text{Cov}(\mathbf{u}, \mathbf{W}\gamma)\\
&= \Lambda^p\mathbf{V}^p + (\mathbf{I} - \Lambda^p)\mathbf{S}^p(\mathbf{I} - \Lambda^p)',
\end{aligned}
$$

where

$$
\mathbf{S} = \mathbf{W}\{\mathbf{W}'[(\mathbf{X}'\Sigma^{-1}\mathbf{X})^{-1} + \mathrm{T}]^{-1}\mathbf{W}\}^{-1}\mathbf{W}'.
$$

In a similar manner, the conditional expectation for T given $\mathbf{y}$, and previous values for Σ and T is

$$E(\Sigma \mathbf{u}_j \mathbf{u}_j') = E[\Sigma(\beta_j - \mathbf{W}_j\gamma)(\beta_j - \mathbf{W}_j\gamma)'] \qquad (3.23)$$
$$= \Sigma \mathbf{u}^p_j \mathbf{u}^p_j' + \Sigma(\Lambda^p_j \mathbf{V}^p_j + \Lambda^p_j \mathbf{S}^p_j \Lambda^p_j').$$

The estimated expected statistics derived from (3.22) and (3.23) are passed to the maximization step to produce restricted maximum likelihood estimates for σ^2 and T respectively (Harville, 1976; and Laird and Ware, 1982). Let us define $\Sigma^* = I\sigma^{2*}$ and T^* to be the estimates that maximize the restricted likelihood. The empirical Bayes approach substitutes these variance-covariance components as if they were known constants. This strategy reduces the information from two distributions into point estimates. Hence, empirical Bayes inferences about γ are based on the following joint posterior distribution:

$$p(\gamma \mid \mathbf{y}, \Sigma=\Sigma^*, T=T^*) = \int p(\mathbf{y} \mid \beta, \Sigma^*)\, p(\beta \mid \gamma, T^*)\, \partial\beta. \qquad (3.24)$$

The empirical Bayes approach does not take into consideration the uncertainty of our knowledge of the unknown variance-covariance matrices Σ and T. The estimates of T in a small sample situation may not be estimated with very much precision. The empirical Bayes approach underestimates the uncertainty associated with T and to a lesser extent Σ because the point estimates for these parameters do not provide any information about their precision. As a result, if other probable values for T produce substantially different values for γ^*, then the dispersion of γ^* will be underestimated. The asymptotically normal properties of maximum likelihood estimators makes this concern less important in large sample problems or in situations where the most probable values of T do not substantially effect the

estimation of γ^{*}.

The Data Augmentation Algorithm

The Bayesian approach provides an alternative to the empirical Bayes approach and explicitly incorporates the uncertainty or precision of the variance-covariance components in the model. The essential difference in these two approaches is that the Bayesian approach specifies a prior distribution for the variance-covariance components in the model, rather than summarizing this information into a point estimate. Hence, Bayesian inferences about γ are base on

$$p(\gamma \mid y, \Sigma, T) = \iiint p(y \mid \beta, \Sigma)\, p(\beta \mid \gamma, T)\, p(\gamma)\, p(\Sigma)\, p(T)\, \partial\beta\, \partial\Sigma\, \partial T. \quad (3.25)$$

This Bayesian approach provides more information about the precision of the standard error of the posterior distribution (Lindley, 1972) than is available with some form of point estimate substitution for the variance-covariance components. In particular, the Bayesian approach presented in (3.25) should provide standard errors for γ that tend to be larger than those in (3.24). This should be especially evident in small sample situations.

The problem with the Bayesian solution to the hierarchical linear model is that even for reasonable priors, it is exceptionally complicated to execute. The required integration is multidimensional, the densities are mathematically complex and cannot all be expressed in closed form (Lindley, 1972). The data augmentation algorithm is an alternative procedure for the implementation of the Bayesian approach that can approximate (3.25) without necessitating the complex integration.

The data augmentation algorithm is a method for calculation of the posterior distribution of a model's parameters without integration. The basic idea behind this algorithm is to supplement the observed data with some latent data, resulting in an augmented data set. In many cases, given this augmented data, the analysis problem, which initially was difficult or intractable, becomes relatively straightforward. In the general hierarchical linear model, each successive level is more difficult to sample because the relevant populations become smaller. For example, it is easier to sample n students than to sample the same number of classes and as a result, the data at higher levels becomes sparse. For the specific hierarchical model under consideration, if large quantities of data were available pertinent to the between-class parameters γ and T, the analysis would be much easier. The data augmentation algorithm developed by Tanner and Wong (1987) provides a strategy to generate the required latent data for the model's between-class parameters to facilitate the statistical analysis.

In order to develop this algorithm, further assumptions are required that were unnecessary for the empirical Bayes procedure. These additional assumptions specify prior distributions for the parameters σ^2 and T. These conjugate prior distributions are

$$\sigma^2 \sim \nu_0\sigma_0^2\chi^{-2}(\nu_0),$$

and

$$T \sim W^{-1}(\Psi, \upsilon).$$

These assumptions were presented previously in (2.12) and (2.13). One

consequence of these assumptions is that the uncertainty resulting from unknown variance-covariance components is included explicitly in the model.

The data augmentation algorithm uses these assumptions in addition to the data to generate a large set of latent or unobserved data by Monte Carlo sampling. If this set of augmented data is sufficiently large, it can be used as a finite approximation to the true posterior distribution. This algorithm is iterative and each successive iteration produces a better approximation, until convergence, to the true posterior distribution. The a priori knowledge assumed for the actual parameter values of the prior distributions are vague or noninformative in their contribution to the posterior distribution. It is the assumptions concerning the model in conjunction with the distribution of the data that determines the joint posterior distribution.

The joint distribution of the model is

$$p(y, \beta, \Sigma, \gamma, T) = \tag{3.26}$$
$$p_1(y \mid \beta, \Sigma)\ p_2(\beta \mid \Sigma, \gamma, T)\ p_3(\Sigma, \gamma, T).$$

This joint distribution can be rewritten in two alternative expressions (Morris, 1987) as

$$q_1(y)\ q_2(\beta, \Sigma \mid y)\ q_3(\gamma, T \mid \beta, \Sigma), \tag{3.27}$$

or

$$r_1(y)\ r_2(\beta, \Sigma \mid y, \gamma, T)\ r_3(\gamma, T \mid y). \tag{3.28}$$

The above two alternative forms contain densities derived from the p_i densities in

(3.26) and $q_1(y) = r_1(y)$. The joint posterior density may be written as

$$p(\beta, \Sigma, \gamma, T \mid y) \propto p(y, \beta, \Sigma, \gamma, T) / q_1(y). \tag{3.29}$$

First, to calculate the joint posterior density presented in (3.29), the knowledge of q_3 would result in the straightforward calculation of the joint posterior density given by r_2 because the joint posterior density of γ and T is known. Or, on the other hand, if r_2 were known then the joint posterior density of q_3 would follow directly because the joint posterior distribution of β and Σ is known. The dependency between q_3 and r_2, in terms of the joint posterior density of the parameters, suggests an iterative solution to this problem, since neither q_3 or r_2 is generally known. The data augmentation algorithm provides an ingenious procedure to overcome this dilemma.

The data augmentation algorithm, as outlined by Tanner and Wong (1987), has two basic steps, the imputation and posterior steps. The first step is the imputation step. This step generates latent data to augment the observed data. Initially, suppose parameters β and Σ from r_2 were observed, then γ^* can be calculated, where the asterisk (*) indicates an estimated mean. Next, given γ^* just calculated and a previously observed T, m new γ's are sampled by Monte Carlo methods from the posterior distribution of γ. This is the point in the algorithm where latent data may be generated to augment the observed data. With the knowledge of the m γ's, m T^*'s can be calculated and subsequently, m T's are sampled from the posterior distribution of T using Monte Carlo methods. The result of the imputation step is m parameters γ and T approximating a sample from q_3.

The second step is the posterior step. This step uses the augmented data resulting from the previous imputation step to obtain a sample of m β's and Σ's. Given m parameters γ and T from q_3 and a previously observed Σ, m values for β^* can be calculated and used as estimated means to sample m updated β's by Monte Carlo methods. Subsequently, m Σ^*'s can be calculated and then m Σ's can be sampled by Monte Carlo methods. The results from this step are m parameters β and Σ approximating a sample from r_2.

Initially, the results from the two steps of the data augmentation algorithm produce rough approximations. To improve the approximations, the results from the posterior step can be considered as an intermediate approximation of r_2 instead of a final one and recycled back to the imputation step to generate a new sample to update q_3. The size of m may be increased in the imputation step during an iteration until it reaches some maximum value. At this point, the value for m may be held constant. This iterative process between the imputation and posterior steps has been shown by Tanner and Wong (1987) to converge in probability to the desired posterior distribution under mild regularity conditions. For large values of m, for example m = 2048, the mixture of the densities produced from the imputation and posterior steps can be considered a finite approximation to the desired joint posterior density presented in (3.29).

One advantage of this algorithm is that not only are point estimates generated as a mixture of densities, but the results also include finite approximations to the true joint posterior distribution. For example, the calculated values for a parameter of interest can be sorted in order and then the $\alpha/2\%$ tails of the distribution can be easily determined. This procedure results in a $(1 - \alpha)\%$ highest

posterior density (HPD) interval for the parameter. Thus, this approach provides an alternative to the sampling theory approach, which is especially useful when the analytically determined sampling distribution of an estimator is not known and the implementation of the appropriate conditional distributions such as r_2 and q_3 do not offer insurmountable problems.

The details for the implementation of the data augmentation algorithm are presented in the following sections. This presentation will begin with a brief discussion concerning the calculation of initial values required to begin the data augmentation algorithm.

Initial Values

Initial values for β, Σ and T are required for the start of this algorithm. Parameter estimates may be calculated by a variety of techniques. One method to calculate these parameter estimates is to use the empirical Bayes procedure. The final estimates from the empirical Bayes procedure can be used to supply initial values to the Bayes approach using the data augmentation algorithm. Since these two statistical procedures are compared in this study, the final estimates of the empirical Bayes procedure were not used to help minimize dependencies between methods. Least squares estimates of β were used as initial starting values for this parameter. The first iteration results from the empirical Bayes procedure for Σ and T were used as starting values because this method constrains estimates of these parameters to their respective parameter spaces.

The Imputation Step

Given a set of m pairs (β, Σ) from $r_2(\beta, \Sigma \mid y, \gamma, T)$, m new values for γ^* and T^* can be calculated. These posterior means are used to generate a sample of size m' of the parameter pairs (γ, T) from q_3, where $m \leq m'$. The first iteration of this step $m = 1$ and may be increase so that $m' = 2$ at the completion of the imputation step. We note that γ is from a normal distribution given by

$$\gamma^{(i)} \sim N(\gamma^{(i)*}, (\mathbf{W}'\mathbf{T}^{-1}\mathbf{W})^{-1}), \tag{3.30}$$

and q_3 may be expressed in the following form:

$$q_3(\gamma, T \mid \beta, \Sigma) = q_{3\cdot}(\gamma \mid T, \beta, \Sigma)\, q_{3\cdot\cdot}(T), \tag{3.31}$$

where $q_{3\cdot\cdot}$ represents the noninformative prior distribution of T. The calculation of γ^* does not depend on T, because T is constant with respect to β. The m values for γ^* can be calculated as follows:

$$\gamma^{*(i)} = \Sigma(\mathbf{W}_j'\mathbf{W}_j)^{-1}\, \Sigma\mathbf{W}_j'\beta_j^{(i)}, \tag{3.32}$$

where

$i = 1, 2, ..., m$, and

$j = 1, 2, ..., k$.

Once the $\gamma^{(i)*}$'s are calculated, a sample is drawn by Monte Carlo sampling from (3.30), the posterior distribution of γ. Let us denote the m T's from the

previous iteration of the imputation step $T_p^{(i)}$. Let

$$\mathbf{A}^{(i)} = [\Sigma(\mathbf{W}_j'\mathbf{T}_p^{(i)-1}\mathbf{W}_j)]^{-1}, \tag{3.33}$$

for

$$i = 1, 2, ..., m,$$

where the $\mathbf{A}^{(i)}$'s provide the current approximation to the dispersion of γ. The $\mathbf{A}^{(i)}$'s are factored such that

$$\mathbf{A}^{(i)} = \mathbf{B}^{(i)}\mathbf{B}^{(i)'}, \tag{3.34}$$

where $\mathbf{B}^{(i)}$ is a lower triangular Cholesky factor of $\mathbf{A}^{(i)}$. The matrix equation used to generate a new $\gamma^{(i)}$ is

$$\gamma^{(l)} = \gamma^{(i)*} + \mathbf{B}^{(i)}\mathbf{x}^{(l)}, \tag{3.35}$$

where

$$i = 1, 2, ..., m,$$

$$l = 1, 2, ..., m' \text{ and } m \leq m'.$$

The vector $\mathbf{x}^{(l)}$ contains elements that are independently and identically distributed N(0, 1). The data may be augmented using (3.35). For example, let m' = 2m. Two $\mathbf{x}$ vectors can be generated for each pair $\gamma^{(i)*}$ and $\mathbf{B}^{(i)}$. Hence, two new γ vectors may be generated.

Once m' new γ's have been generated, this information can be used to

update the posterior distribution of T. Let

$$T^{(i)*} = (1/k)\ C^{(i)}, \tag{3.36}$$

where

$C^{(i)} = \Sigma(\beta_j^{(i)} - W_j\gamma^{(i)})(\beta_j^{(i)} - W_j\gamma^{(i)})'$, and now

$i = 1, 2, ..., m'$.

If the i^{th} sample covariance matrix is $T^{(i)*}$, where $T^{(i)*}$ has the Wishart distribution with parameters T and k, and T has the a priori inverse Wishart distribution $W^{-1}(\Psi, \upsilon)$, then the conditional distribution of T (given $T^{(i)*}$) is $W^{-1}(T^{(i)*} + \Psi, k + \upsilon)$ (Anderson, 1984). Thus, if we assume the precision matrix Ψ approaches **0** and its degrees of freedom parameter υ approaches zero to indicate a noninformative prior, then the posterior distribution of T is

$$T \sim W^{-1}(T^{(i)*}, k). \tag{3.37}$$

Jones (1985), and Smith and Hocking (1972) describe a method that uses Bartlett's (1933) decomposition along with $T^{(i)*}$ to obtain a random sample from the distribution in (3.37).

To obtain a sample of $T^{(i)}$'s from the distribution given in (3.37) we find the Cholesky factor of the $T^{(i)*}$'s such that $T^{(i)*} = \mathbf{D}^{(i)}\mathbf{D}'^{(i)}$, where $\mathbf{D}^{(i)}$ is a lower triangular matrix. For $\mathbf{E}^{(i)}$, a lower triangular matrix, let

$$\mathbf{F}^{(i)} = \mathbf{D}^{(i)}\mathbf{E}^{(i)}\mathbf{E}'^{(i)}\mathbf{D}'^{(i)} = (\mathbf{D}^{(i)}\mathbf{E}^{(i)})(\mathbf{D}^{(i)}\mathbf{E}^{(i)})'. \tag{3.38}$$

Furthermore, if e_{ij} is an element of **E**, and e^2_{jj} is an independent chi-square variable with (k - j + 1) degrees of freedom and the elements below the diagonal are independently distributed N(0, 1), then $\mathbf{F}^{(i)}$ will have the required inverse Wishart distribution with parameters $T^{(i)*}$ and k (Jones, 1985; and Anderson, 1984).

Now, for each $T^{(i)*}$ calculated in (3.36) an updated $T^{(i)}$ is sampled from (3.37). This is accomplished by generating $\mathbf{F}^{(i)}$ from (3.38) using Monte Carlo methods and setting $T^{(i)} = 1/k\ \mathbf{F}^{(i)}$. These new values for $\gamma^{(i)}$ and $T^{(i)}$ are passed to the posterior step to generate new updated values for β and Σ.

The Posterior Step

The data generated from q_3 by the execution of the I-Step consists of the pairs $(\gamma^{(i)}, T^{(i)})$, for i = 1, 2, ..., m. These values are considered known during the execution of the P-Step. Given these data, the posterior step calculates m β^*'s and Σ^*'s. This information is used to realize a sample of size m from r_2. The desired posterior density r_2 can be rewritten as

$$r_2(\beta, \Sigma \mid \mathbf{y}, \gamma, T) = r_{2'}(\beta \mid \mathbf{y}, \Sigma, \gamma, T)\ r_{2''}(\Sigma). \tag{3.39}$$

We note that $r_{2''}$ is the noninformative prior assumed for Σ.

Let $\Sigma_p^{(i)}$ indicate the previous iterations sample from the posterior distribution of Σ. New values for the β^*'s can be calculated with the following equation:

$$\beta_j^{(i)} = \Lambda_j^{(i)}\beta_j^0 + (\mathbf{I} - \Lambda_j^{(i)})\mathbf{W}_j\gamma^{(i)}, \tag{3.40}$$

where

$$\Lambda_j^{(i)} = (\mathbf{V}_j^{(i)-1} + \mathbf{T}^{(i)-1})^{-1}\mathbf{V}_j^{(i)-1},$$

and

$$\mathbf{V}_j^{(i)-1} = (\mathbf{X}_j'\Sigma_p^{(i)-1}\mathbf{X}_j)^{-1}.$$

A new set of β's can be sampled from the following distribution:

$$\beta_j^{(i)} \sim N(\beta_j^{(i)*}, D(\beta_j^{(i)}), \tag{3.41}$$

where

$$D(\beta_j^{(i)}) = (\mathbf{V}_j^{(i)-1}) + \mathbf{T}^{(i)-1})^{-1}.$$

Samples are drawn from (3.41), an approximation of r_2, in much the same manner as the $\gamma^{(i)}$'s were sampled in the imputation step. First, the matrix $D(\beta_j^{(i)})$ is factored by the Cholesky method such that

$$D(\beta_j^{(i)}) = \mathbf{L}_j^{(i)}\mathbf{L}_j^{(i)'}, \tag{3.42}$$

where $\mathbf{L}_j^{(i)}$ is a lower triangular matrix. Next, the $\beta_j^{(i)}$'s are sampled with the following equation:

$$\beta_j^{(i)} = \beta_j^{(i)*} + \mathbf{L}_j^{(i)}\mathbf{x}_j^{(i)}, \tag{3.43}$$

where $\mathbf{x}_j^{(i)}$ is a vector containing independent and identically distributed elements sampled from N(0, 1).

After generating new $\beta_j^{(i)}$'s from (3.43), the $\Sigma^{(i)*} = \sigma^{2(i)*}\mathbf{I}$ can be updated to

reflect these new values. Let

$$\sigma^{2(i)*} = 1/N \sum(y_j - X_j\beta_j^{(i)})'(y_j - X_j\beta_j^{(i)}), \tag{3.44}$$

where

$$N = \sum n_j.$$

It has been assumed that $\sigma^{2(i)*}$ has a chi-square distribution and σ_0^2 has an inverse chi-square prior distribution. This prior distribution of σ_0^2 is noninformative and may be considered a very small constant in its contribution to the posterior distribution of the $\sigma^{2(i)}$'s when the degrees of freedom for σ_0^2 approach 0 (Lindley, 1965b). The updated posterior values of $\sigma^{2(i)}$ can be generated by Monte Carlo methods as follows:

$$\sigma^{2(i)} = N\sigma^{2(i)*} / \chi^{2(i)}(\nu), \tag{3.45}$$

where the $\chi^{2(i)}(\nu)$ are independent chi-square variates with the degrees of freedom parameter $\nu = N$.

This simulated sample from r_2 of m parameter pairs $(\beta^{(i)}, \Sigma^{(i)})$ are passed back to the imputation step to begin a new iteration.

Some Additional Notes on the Data Augmentation Algorithm.

To begin the first iteration, there is only one estimate of β, γ, and T available for the first iteration of the imputation step. The parameters γ and T are augmented in this step of the algorithm. Each execution of the imputation step

produces twice as many samples from the posterior density as the number entering this step. This doubling of the latent data continues up to some predetermined maximum. The rate the data is augmented is a variable of the algorithm. Doubling the data in the imputation step was chosen as a slow linear rate so that the number of iterations required for convergence would not be as reliant on the initial values as they would if m was set at its maximum value right at the beginning. After this maximum value has been reached, it is held constant until a stopping criterion for the algorithm is reached.

Traditionally, an iterative process can be said to converge when some function based on the difference between the previous and the current iteration values is less than some specified tolerance. The implementation of the data augmentation algorithm includes the introduction of random elements, and as a result the convergence criterion for this algorithm will be quite different than the one used for the EM algorithm. The data augmentation algorithm will produce a stochastic trend across iterations rather than a strictly increasing monotonic function evaluated with the EM algorithm. Since this is a simulation study, the population values of all the parameters are known. These population values can be compared to the values obtained from the posterior distributions calculated with the data augmentation algorithm at selected iterations. This information obtained from a number of different iterations across experimental data sets will be examined to provide some insight into an appropriate number of iterations to provide the desired level of convergence. Determining convergence is a major unsolved problem of the data augmentation algorithm.

CHAPTER IV

METHOD

Research Questions

The primary questions of interest are introduced in this section. The specific methodological details for investigating these questions are presented in subsequent sections.

Determining a Convergence Criterion

The iterative data augmentation algorithm lacks a strategy for determining convergence. The convergence of this algorithm is dependent on a number of factors, such as, the model, the assumptions, the data, the number of augmented data sets that are generated, and the number of iterations. How do we know when this algorithm has converged?

Comparing the Accuracy of the Two Statistical Procedures

There is little known about the performance of the empirical Bayes approach using the EM algorithm in small sample hierarchical problems and the Bayesian approach using the data augmentation algorithm is virtually untested as an applied statistical technique. In many studies, the most important parameter is the γ vector. In addition, the variance-covariance matrices Σ and T are important parameters to

accurately recover because they provide information on the dispersion of β and γ. Which of the two competing statistical procedures recover the parameters of the model from small sample hierarchical data with the greatest accuracy?

The Error Rates

There are two possible types of errors that can be committed when making an inference in a hypothesis testing situation. A type I error is made when the null hypothesis is rejected when it is in fact true. Traditionally, researchers have tried to control the type I error rate, α. The origins of the 0.05 level of statistical significance can be traced back to Sir Ronald Fisher in 1925 (Cowles and Davis, 1982) and a narrow range of values around the 0.05 level have persisted. Under the conditions of this simulation, are the observed type I error rates equivalent to the nominal error rates for the empirical Bayes and the Bayes statistical procedures?

The other error that can be made is when the null hypothesis is not rejected when in fact the alternative is true. This is referred to as a type II error. The type II error rate, β, is used to define the statistical power of a test, $(1 - \beta)$. This translates into a statement about the probability of detecting a true alternative hypothesis. Is the statistical power for both procedures equivalent? If the statistical power is different, which procedure is the most powerful?

The Computational Efficiency of the Statistical Algorithms

The computational resources required by the data augmentation algorithm are unknown. This procedure generates large quantities of data and it is also iterative. Consequently, it has the potential to consume substantial amounts of computer

resources. Are the computational resources required by the data augmentation algorithm within practical limits for the application of this method to substantive problems in education? How do the two statistical procedures compare in terms of the computer resources they require?

The EM algorithm is also an iterative algorithm. How many iterations are required for convergence? Do the required number of iterations change when the EM algorithm is applied to data from different populations? In particular, does the number of iterations increase as the conditional variance of τ_{11} decreases?

The Design and Sampling Plan for the Simulation

The primary design factor in this simulation study is the size of the between-class parameter γ_{11}. The empirical Bayes approach using the EM algorithm and the Bayesian approach using the data augmentation algorithm are compared under three experimental conditions. The γ_{11} parameter varies across the three experimental conditions. The relative magnitude of this interaction (ATI) effect parameter is essentially determined by ρ_1, the correlation of β_{j1} with W_j, where W_j is the between-class predictor.

Five hundred replications of an experiment are generated for each experimental condition. This number was chosen as a compromise between the number of replications necessary to make an informed inference about the statistical results produced by the two procedures and the required computer processing time. Computer processing time is a concern because both algorithms are iterative and computation intensive. This is especially true for the data augmentation algorithm.

The number of students assigned to each class in this simulation is drawn

randomly from a distribution of class size. This distribution models a plausible distribution of elementary school class size. The expected class size from this distribution is 24.03. Table 4-1 displays the distribution of class size.

Table 4-1

Class Size Frequency Distribution

Class Size	Percentage
30	0.005
29	0.030
28	0.060
27	0.110
26	0.125
25	0.135
24	0.130
23	0.125
22	0.095
21	0.070
20	0.050
19	0.030
18	0.020
17	0.010
16	0.005

Creating the Data

A number of issues must be addressed before the appropriate data can be generated. The parameters necessary to generate the data must be defined and then suitable values must be selected before the actual data sets can be constructed. Another consideration is the technique used to generate the required random numbers and transforming them to the appropriate distributions. These issues are covered next.

Parameters of the Model

For the specific hierarchical linear model investigated, the predictor variable, x_{ij1}, at the within-class level and the error term, r_{ij}, are identically and independently distributed normally with means equal to zero and variances equal to one. Furthermore, the variables x_{ij1}, r_{ij}, u_{j0}, and u_{j1} are mutually independent. Finally, T is assumed to be a block diagonal matrix with homogeneous diagonal submatrices. This assumption permits generation of a common T_j matrix and as a consequence a pooled estimate of this variance-covariance matrix is appropriate.

To generate data with the required distribution, we must define some additional parameters. The method used to generate the necessary standardized hierarchical data was developed by Bassiri (1988) and modified to fit this specific model. This special case based on standardized hierarchical data changes the location and scale but it does not affect generalization to other normal distribution parameter values. Let the pooled within-class parameters be defined in the following manner:

γ_1 is the pooled slope,

ρ_{xy} is the pooled within-class and within-treatment correlation of x with y,

σ_y^2 is the unconditional pooled variance of y, and

σ_x^2 is the pooled variance of x.

Recall that τ_{00} and τ_{11} are the diagonal elements of T_j corresponding to the $Var(u_{j0})$ and $Var(u_{j1})$ respectively. Also, $Cov(u_{j0}, u_{j1}) = 0$ which implies that $\tau_{01} = \tau_{10} = 0$. The following is a list of required definitions.

1. $c = \tau_{11} / \tau_{00}$ and $\tau_{11} = c\tau_{00}$. (4.01)

2. $\gamma_1^2 = \rho_{xy}^2 \sigma_y^2 / \sigma_x^2$, (4.02)

and

$$\sigma_y^2 = \mathrm{Var}(x_{ij1}\gamma_1 + x_{ij1}u_{j1} + r_{ij}) \quad (4.03)$$
$$= \gamma_1^2 + \tau_{11} + \mathrm{Var}(r_{ij})$$
$$= \gamma_1^2 + c\tau_{00} + 1.0.$$

The results of (4.03) are substituted into (4.02) and the expression for γ_1^2 can be rewritten as

$$\gamma_1^2 = \rho_{xy}^2(\gamma_1^2 + c\tau_{00} + 1.0) \quad (4.04)$$
$$= (\rho_{xy}^2 / (1.0 - \rho_{xy}^2))\ (c\tau_{00} + 1.0).$$

3. $d = \tau_{00} / (\tau_{00} + \sigma_y^2)$, (4.05)

where d is the intraclass correlation. Next, the expression for the pooled within-class variance of y can be expressed in a new form by substituting (4.04) into (4.03) and

$$\sigma_y^2 = \{(\rho_{xy}^2 / (1.0 - \rho_x^2))\ (c\tau_{00} + 1.0)\} + c\tau_{00} + 1.0. \quad (4.06)$$

Substituting equation (4.06) for σ_y^2 into equation (4.05) and solving for τ_{00} results in

$$\tau_{00} = d / ((1.0 - d)(1.0 - \rho_{xy}^2) - cd). \quad (4.07)$$

The values for c are constrained by $0 < c < (1 - \rho_{xy}^2)(1 - d) / d$, so that the variance component τ_{00} is greater than zero.

Equation (4.07) implies that once values are selected for the parameters d, ρ_{xy}, and c, the value for the variance component τ_{00} is determined. From the definition of c in (4.01), it can be seen that τ_{11} is also determined.

Definitions for the conditional variances τ_{00} and τ_{11} given **W** must be considered in addition to the corresponding unconditional variance terms defined previously. Let ρ_0 be the correlation between β_{j0} and W_j and ρ_1 is the correlation between β_{j1} and W_j, for W_j as defined in (2.02a) and (2.02b). Let

$$\gamma_{01} = \rho_0(\tau_{00})^{1/2}, \tag{4.08}$$

and

$$\gamma_{11} = \rho_1(\tau_{11})^{1/2}. \tag{4.09}$$

The conditional variances for τ_{00} and τ_{11} given **W** are defined as follows,

$$\begin{aligned} \tau_{00|W} &= \tau_{00} - \gamma_{00}^2 \\ &= \tau_{00}(1.0 - \rho_0^2), \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} \tau_{11|W} &= \tau_{11} - \gamma_{11}^2 \\ &= \tau_{11}(1.0 - \rho_1^2). \end{aligned} \tag{4.11}$$

The covariance terms τ_{01} and τ_{10} are set to zero. The two between-class parameters γ_{01} and γ_{11} are determined by the correlation coefficients ρ_0 and ρ_1. Since ρ_0 is set to zero, γ_{01} will also be equal to zero. In addition, γ_{00} is equal to zero because this is the central value for a standardized hierarchical linear model. And finally, γ_{10} reflects the base for the within-class slopes.

The Parameter Values Selected

The number of classes was selected at ten to represent a small but realistic number of classes that might typically arise in an experimental research project at the class level. This small sample size of classes was selected because it is expected that the difference between the two statistical procedures will be greatest under this condition.

The three levels of the correlation coefficient ρ_1 determining the relative size of γ_{11} are 0.0, 0.3 and 0.6. The value 0.0 was selected for the type I error rate analysis. The other two values were selected to investigate the power of the test over a range of plausible values for the between-class correlation coefficient related to γ_{11}. Also, the two algorithms may recover T differentially as the correlation ρ_1 increases in size. Table 4-2 illustrates the three experimental data conditions and the values selected for ρ_1.

Table 4-2

The Design for the Experimental Conditions

	Experimental Conditions		
Correlations	1	2	3
ρ_1	0.0	0.3	0.6

Many student characteristics such as motivation, aptitudes, and interest have a moderate linear relationship with achievement if the measures are chosen judiciously. Since there is only one independent variable at the within-class level, it

is assumed that this single independent measure was chosen with care. To reflect this situation, the actual value of the correlation coefficient for ρ_{xy} was set at 0.60.

The intraclass correlation, d, is generally small in educational research. Kish (1965) reports that most intraclass correlations are less than 0.15. Raudenbush and Bryk (1986) estimated the intraclass correlation to be 0.177 in a national high school educational research project. Blair and Higgins (1986) picked a value of 0.20 for the intraclass correlation used in their simulation study. There is some evidence that the intraclass correlation associated with educational data has a rather narrow range of values (Blair et al., 1983). The intraclass correlation for this simulation was selected to be 0.18.

Research findings do not routinely report the value of c, the ratio τ_{11} / τ_{00}. Thus, choosing a likely value for c is not as straightforward. There is some indication that the value of c in an educational context is rather small. Bryk and Raudenbush (1987) found values of c equal to 0.11 and 0.05, and Raudenbush and Bryk (1986) found a value of 0.10. A value of 0.10 was selected for this parameter.

The results determined by these selected parameter values are summarized in Table 4-3.

Table 4-3

Parameter Values for the Three Experimental Conditions

	Experimental Conditions		
Parameters	1	2	3
ρ_1	0.0	0.3	0.6
γ_{00}	0.0	0.0	0.0
γ_{01}	0.0	0.0	0.0
γ_{10}	0.7632	0.7632	0.7632
γ_{11}	0.0000	0.0565	0.1130
σ^2	1.0	1.0	1.0
$\tau_{00\|w}$	0.3552	0.3552	0.3552
$\tau_{11\|w}$	0.0355	0.0323	0.0227
$\tau_{01\|w} = \tau_{10\|w}$	0.0	0.0	0.0

The Steps for the Data Generation Procedure

The basic strategy for data generation is to create sufficient statistics for each class and then collect k classes of sufficient statistics into a complete set of data simulating one experiment. There are a couple of reasons for this approach over the more traditional method of generating data specifically for each unit of statistical analysis. The first reason is that both the HLM computer program developed by Bryk et al. (1986) and the data augmentation program developed for this study operate on a set of sufficient statistics as input. The other reason is to reduce the number of observations that must be created and manipulated. The following data generation procedure was originally presented by Bassiri (1988) with some modifications incorporated to fit the unique requirements of this study.

Let i = 1, 2, ..., n_j index students within the j^{th} class. The five sufficient statistics required for each class are

$\sum x_i$, $\sum x_i^2$, $\sum r_i$, $\sum r_i^2$, and $\sum x_i r_i$.

The within-class data, the x_{ij1}'s and r_{ij}'s, are independently and identically distributed N(0, 1), which implies that the $Cov(x_{ij1}, r_{ij}) = 0$.

The following steps are used to simulate the data for one class and k repetitions constitute one simulated experiment. These steps are repeated 500 times for each experimental condition depicted in Table 4-2. The only parameter value of the data that changed was ρ_1. Let $i = 1, 2, ..., n_j$ indicate students within the j^{th} class and the required steps are listed next.

1. Generate a value for n_j sampled randomly from the distribution of class size presented in Table 4-2.
2. Generate

$$\sum x_i \sim N(0, n_j). \tag{4.12}$$

3. Generate

$$\sum (x_i - \bar{x}.)^2 \sim \chi^2_{(n_j-1)}, \tag{4.13}$$

and then use the results from step 2 to compute

$$\sum x_i^2 = \sum (x_i - \bar{x}.)^2 + (\sum x_i)^2 / n_j. \tag{4.14}$$

4. Generate

$$\sum r_i \sim N(0, n_j). \tag{4.15}$$

5. Similar to step 3, generate

$$\sum (r_i - \bar{r}.)^2 \sim \chi^2_{(n_j-1)}, \tag{4.16}$$

and then using the results from step 4 to compute

$$\sum r_i^2 = \sum (r_i - \bar{r}.)^2 + (\sum r_i)^2 / n_j. \quad (4.17)$$

6. First generate $z \sim N(0, 1)$ and then construct a t variable with (n_j-2) degrees of freedom as follows:

$$t = z / (\chi_{(n_j-2)} / (n_j - 2))^{1/2}, \quad (4.18)$$

then compute the sample correlation coefficient R

$$R = t / (t^2 + n_j - 2)^{1/2}, \quad (4.19)$$

and

$$\sum x_i r_i = R(\sum (x_i - \bar{x}.)^2)^{1/2} (\sum (r_i - \bar{r}.)^2)^{1/2} + (\sum x_i)(\sum r_i)/n_j. \quad (4.20)$$

7. Half the classes are randomly assigned to the experimental condition and the remaining classes to the control group.

8. Recall that γ_{00} has been set to zero and γ_{10} equals the pooled within-class regression coefficient. The correlation coefficient, ρ_1, is set according to Table 4-2 and ρ_0 equals zero. The conditional variance terms, $\tau_{00|w}$ and $\tau_{11|w}$ are calculated from (4.10) and (4.11) respectively. When γ_{01} or γ_{11} are set to zero, their conditional and unconditional variance terms are equivalent. Since the $Cov(u_{j0}, u_{j1}) = 0$, the elements of $\mathbf{u}_j$ can be generated independently from the following distributions:

$$u_{0j} \sim N(0, \tau_{00|w}), \quad (4.21)$$

and

$$u_{1j} \sim N(0, \tau_{11|w}). \quad (4.22)$$

9. Recall from (2.02a) and (2.02b) that

$$\beta_{j0} = \gamma_{00} + \gamma_{01} W_j + u_{j0},$$

and

$$\beta_{j1} = \gamma_{10} + \gamma_{01} W_j + u_{j1},$$

where γ_{00} is set to zero, γ_{01} and γ_{11} are determined from equations (4.08) and (4.09) respectively, and γ_{10} is computed by taking the square root of equation (4.04).

10. The last step in the data generation procedure calculates $\Sigma\, y_{ij}$, $\Sigma\, y_{ij}^2$, and $\Sigma\, x_{ij1} y_{ij}$ from the sufficient statistics already available and the values selected for the parameters. The three desired quantities are calculated as follows:

$$\Sigma\, y_{ij} = \Sigma\, (\beta_{j0} + \beta_{j1} x_{ij1} + r_{ij}) \tag{4.23}$$

$$= n_j \beta_{j0} + \beta_{j1} \Sigma\, x_{ij1} + \Sigma\, r_{ij},$$

$$\Sigma\, y_{ij}^2 = \Sigma\, (\beta_{j0} + \beta_{j1} x_{ij1} + r_{ij})^2 \tag{4.24}$$

$$= n_j \beta_{j0}^2 + \beta_{j1}^2 \Sigma\, x_{ij1}^2 + \Sigma\, r_{ij}^2 + 2\beta_{j0}\beta_{j1} \Sigma\, x_{ij1} + 2\beta_{j0} \Sigma\, r_{ij}$$

$$+ 2\beta_{j1} \Sigma\, x_{ij1} r_{ij},$$

and

$$\Sigma\, x_{ij1} y_{ij} = \beta_{j0} \Sigma\, x_{ij1} + \beta_{j1} \Sigma\, x_{ij1}^2 + \Sigma\, x_{ij1} r_{ij}. \tag{4.25}$$

Random Number Generation

The random number generator was based on the linear congruential method. This method produces a sequence of random uniform values between 0 and 1. These uniform random values are then transformed from a uniform distribution to another distribution, for example a normal or chi-square distribution. The general form of the linear congruential sequence is given by:

$$X_{i+1} = (aX_i + b) \bmod m, \qquad \text{for } i \geq 0, \tag{4.26}$$

where

m is the modulus,

a is the multiplier,

b is the additive increment, and

X_0 is the initial value or seed.

The two subroutines DRNNOR and DRNCHI from the International Mathematical and Statistical Library (IMSL), Version 10.0 were used to generate and transform random uniform values on the interval (0, 1) into random standard normal and chi-square variables respectively.

The random standard normal deviates produced by the IMSL subroutine DRNNOR were transformed when necessary to another normal distribution with new location and scale parameter values using the following general transformation:

$$Y_i = \mu + \sigma_y x_i, \tag{4.27}$$

where

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

and

$Y_i \sim N(\mu, \sigma_y^2)$.

Evaluating Convergence

To determine a convergence criterion, the first 100 experiments under the no effects condition for γ_{11} are analyzed. The convergence of this algorithm is

investigated with the maximum number of augmented data sets set at 2048. The results from the data augmentation algorithm are evaluated at selected iterations relative to the population parameter values. The mean squared errors resulting from the discrepancy between the results from the data augmentation algorithm and the population parameter values are calculated at 20, 25, 30, 40, 50 and 60 iterations. The calculation of the mean squared errors provide a measure of how effectively the data augmentation algorithm recovers information from the data. The most important parameters to investigate are the variance-covariance components because they are used as weights to calculate the posterior β. In addition, these variance-covariance components are required for inferences about β and γ.

Convergence in the context of the data augmentation algorithm is a stochastic process, since random sampling is an integral element of the algorithm. Thus, convergence must be evaluated over iterations for trends in addition to the usual convergence criterion that tests some function of successive values for inclusion in an arbitrarily small neighborhood. The necessary number of iterations is primarily determined from the stability of the mean squared error terms of σ^2 and T from this first run of 100 experiments. Of course, the mean squared errors of the γ elements are also examined. The required number of iterations is then used for subsequent data augmentation runs.

This "empirical" procedure for determining convergence is used because the properties of the data are known whereas, little is known about the convergence properties of the data augmentation algorithm. This algorithm's convergence is stochastic rather than absolute and this makes it more difficult to determine a criterion for convergence because any function measuring a difference between

iterations will contain random sampling fluctuations. A real time convergence criterion must take into consideration information generated by the algorithm on the parameters of the model over a number of iterations. Generalizing a real time convergence criterion is an important and complex problem that must be addressed in future studies before this method for implementing a Bayesian solution can be used by researchers.

Another related concern is how dependent are the results of the data augmentation algorithm to specific staring values. In this study, the data augmentation algorithm uses the first iteration's estimates of the variance-covariance components from the EM algorithm implemented in the HLM computer program (Bryk, et al., 1986). Two trial replications are run with both good and poor variance-covariance component starting values. The standard good starting values are from the first iteration of the EM algorithm. The poor starting values are determined by multiplying the good starting estimates for σ^2 by 1/2 and T by 2. The results from the data augmentation algorithm are displayed for each of these conditions. In addition, the results from the EM algorithm and the population values for these parameters are displayed.

Evaluating the Accuracy of the Two Statistical Procedures

Descriptive information is calculated and displayed for the important parameters. The parameters γ, σ^2, and T are of principal interest. From the Bayesian approach the posterior distributions of γ, σ^2, and T are available for analysis and their posterior means are easily determined. The empirical Bayes approach provides posterior modal estimates for γ, σ^2, and T. The point estimates

for these parameters are collected from each procedure and averaged over the 500 replications in each experimental condition.

In the unique circumstance of a simulation study, the value of each parameter in the model is known. Therefore, the performance of each statistical procedure may be evaluated with respect to the accuracy they recover the population parameters. The mean squared errors are calculated for each parameter estimate relative its associated population value. To evaluate the relative accuracy these procedures recover the population parameters, the ratio of mean squared errors of the data augmentation algorithm and the EM algorithm are compared for γ, σ^2, and T. These ratios produce F test statistics and are compared to a tabled value with the appropriate degrees of freedom. These tests are performed for each experimental data condition.

Evaluating Type I Error Rates and Power

The type I errors produced from standard two-tailed z-tests and t-tests from the empirical Bayes approach are tabulated for the no effect hypothesis tests for the three parameters γ_{00}, γ_{01} and γ_{11}. The investigation of type I errors for γ_{11} are appropriate only under the first experimental condition. The data augmentation procedure produces a finite approximation to the posterior distribution of γ. An alternative test of a null hypothesis, available with the Bayesian approach, may be determined using the $\alpha/2$ and $(1 - \alpha/2)$ percentiles of a parameters posterior distribution. Since the β_{j1}'s have been set significantly greater than zero, it is inappropriate to include γ_{10} in the type I error analysis. The three tests are evaluated at three levels of significance, 0.01, 0.05 and 0.10. An observed type I

error rate, ρ, is compared to the nominal rate, ρ_0, using a standard test for a binomial proportion. The standardized test statistic is

$$Z = (\rho - \rho_0) / (\rho_0 (1 - \rho_0) / n)^{1/2} \qquad (4.28)$$

This is a large sample normal approximation to the binomial distribution. As a rule of thumb, (Johnson and Bhattacharyya, 1977, p. 203) this approximation is satisfactory when $n\rho$ is greater than 15. Thus, this approximation might be marginal for tests at $\alpha = 0.01$, but it should be satisfactory for tests at the other two levels of significance.

The type I error rates for each procedure are compared using McNemar's test. The following details for McNemar's test are from Fleiss (1981). The appropriate layout for the data is presented next in Table 4-4.

Table 4-4

Data Table for Error Analysis

	Test 1		
Test 2	Correct Decision	Wrong Decision	Total
Correct Decision	a	b	a + b
Wrong Decision	c	d	c + d
Total	a + c	b + d	n

The entries in Table 4-4 represent the number of response pairs that fall into a particular cell. This test for matched pairs with a dichotomous outcome tests the

hypothesis that two proportions are equal. The proportion of test 1 results that correctly fail to reject the null hypothesis is

$$p_1 = (a + c) / n, \tag{4.29}$$

and the proportion of test 2 results that correctly fail to reject the null hypothesis is

$$p_2 = (a + b) / n. \tag{4.30}$$

The difference between the two proportions, $p_2 - p_1$, may be tested with the following statistic with correction for continuity developed by Edwards (1948):

$$\chi^2 = (|b - c| - 1)^2 / (b + c). \tag{4.31}$$

The value of the χ^2 statistic is compared to the central chi-square distribution with one degree of freedom. A large value of the test statistic implies a difference in error rates.

Guarding against type I errors is not the sole concern in hypothesis testing situations. Researchers are particularly interested in the power of a statistical test for a given set of experimental conditions. The more powerful a test for a given level of significance the more sensitive it is in detecting true experimental effects. Power is often measured over a range of possible effect sizes represented by true alternative hypotheses for a predetermined level of significance. In particular, the power of the tests for the true effects represented by γ_{01} and γ_{11} are investigated

with $\alpha = 0.05$. The power of the test is evaluated in a manner similar to the type I error rate. The data are collected and tabulated in the format presented in Table 4-4. The essential difference between type I error analysis and power analysis is in what hypothesis is true. For the type I error analysis the no effect null hypothesis is true and the correct decision is to not reject the null hypothesis. For a power analysis the alternative hypothesis is true and the correct decision is to reject the null hypothesis.

The power of McNemar's test under the conditions of this study is an important consideration. In general, statistical power is related to sample size. Even with the number of experiments set at 500 for each data condition the effective sample size is substantially less than this figure, because McNemar's test statistic only uses the entries in the cells b and c from Table 4-4. The sum of these off diagonal elements represent the total number of divergent responses.

An approximation of the power function for McNemar's test can be used to estimate the power of this test. The following approximations to McNemar's test without the continuity correction is from Miettinen (1968). The approximate power of McNemar's test depends on a discrepancy parameter defined as

$$\delta = E(p_1) - E(p_2). \tag{4.32}$$

Also, the nuisance parameter ζ is defined as

$$\zeta = E(p_1) + E(p_2). \tag{4.33}$$

Miettinen (1968) suggests setting the nuisance parameter ζ at the upper bound of the following expression:

$$|\delta| \leq \psi \leq E(p_1) + E(p_2) - 2E(p_1p_2). \tag{4.34}$$

For example, suppose $E(p_1) = 0.97$ and $E(p_2) = 0.93$, then plausible values for δ and ψ from a type I error analysis between the Bayesian and empirical Bayes approaches with the nominal $\alpha = 0.05$ are

$$\delta = 0.97 - 0.93 = 0.04, \tag{4.35}$$

and

$$\psi = 0.97 + 0.93 - 2(0.97)(0.93) = 0.0958. \tag{4.36}$$

An approximation to the unconditional power function is

$$(1 - \beta) \approx \Phi\{[-\upsilon_{\alpha/2}\psi + (n\psi)^{1/2}|\delta|] / [\psi^2 - (0.25)\delta^2(3 + \psi)]^{1/2}\}, \tag{4.37}$$

where $\upsilon_{\alpha/2}$ is the $100(1 - \alpha/2)$ percentile of the standard normal distribution, n is the total sample size, and Φ denotes the distribution function of a standard normal variate. Accordingly, the power of the test to differentiate between type I error rates of the two statistical procedures is approximately 0.84. This should provide satisfactory power if the results are similar to the those projected.

Evaluating the Implementation of the Data Augmentation Algorithm

The amount of computer resources used by the data augmentation is an important consideration for its application to substantive problems. The amount of central processing unit time will be collected and a mean processing time will be determined for this statistical procedure for each of the simulation conditions. This will provide a per job estimate for solving similar problems on an IBM 3090-180. The EM algorithm implemented in the HLM computer program (Bryk, et al., 1986) has been used extensively in applied research and the amount of computer resources it requires is well within practical limits set for most applied research. The central processing unit time required by the HLM computer program is collected to provide a relative measure of computational efficiency.

The EM algorithm is an iterative statistical procedure and it is of interest to tabulate the number of iterations required for convergence. The number of iterations required to obtain the convergence criterion for each experimental data condition are tabulated to investigate the question of whether of not the parameter values of the data sets effect the number of iterations necessary.

CHAPTER V

SIMULATION RESULTS

The results from the simulation are presented in this chapter in the same order the questions were stated in the preceding chapter. Abbreviations are used in some instances to indicate the statistical procedures. "DA" indicates the Bayesian approach using the data augmentation algorithm and "EM" indicates the empirical Bayes approach using the EM algorithm.

Convergence Criterion Results

A stopping rule was necessary to effectively implement the data augmentation algorithm. This algorithm was developed with the maximum number of augmented data sets fixed at m = 2048. It required eleven iterations to reach this number of augmented data sets. The tests for convergence were carried out on one hundred replications of the first experimental condition. The only element of γ set different than zero was γ_{10}. The primary method used to evaluate convergence was to investigate the mean squared errors between the results of the data augmentation algorithm and the population values of the parameters. The mean squared error terms were calculated at a selected number of iterations for the parameters of interest. These results are presented next in Table 5-1 and Table 5-2.

Table 5-1

MSE Between DA Posterior Means and Population Values for γ

Number of Iterations	Parameters			
	γ_{00}	γ_{01}	γ_{10}	γ_{11}
20	0.035461	0.036138	0.009657	0.007949
25	0.035394	0.036193	0.009592	0.007926
30	0.035258	0.036327	0.009579	0.007950
40	0.035311	0.036220	0.009620	0.007976
50	0.035852	0.036180	0.009539	0.008101
60	0.035460	0.036082	0.009559	0.008017

Based on 100 replications.

Table 5-2

MSE Between DA Posterior Means and Population Values for σ^2 and T

Number of Iterations	Parameters			
	σ^2	τ_{00}	τ_{01}	τ_{11}
20	0.011246	0.036207	0.003193	0.001158
25	0.011269	0.036035	0.003127	0.001143
30	0.011150	0.036734	0.003128	0.001169
40	0.011219	0.036201	0.003135	0.001145
50	0.011141	0.035999	0.003099	0.001113
60	0.011174	0.036785	0.003166	0.001170

Based on 100 replications.

It was expected that the data augmentation algorithm would exhibit convergence with some variance about a fixed point from iteration to iteration. If the algorithm did not converge for a given augmented data set size there would be

a trend in the mean squared error terms in the two tables with these terms generally decreasing as the iterations increased. Inspection of these tables provides no evidence of a general decreasing linear trend for the parameter estimates over the sampled iterations. These tables provide no indication that one value for the number of iterations was superior to another. Since the data augmentation algorithm is very computationally intensive, a smaller number of iterations is desirable. In addition, the initial values for the variance-covariance components for the data augmentation algorithm were the results from the first iteration of the EM algorithm. These initial estimates should provide good starting values. The number of iterations was selected at 25. This was a slightly more conservative criterion than the minimum number of iterations tested, which was 20.

To examine the influence of different initial values, the data augmentation method was run using good and then poor initial values for the variance-covariance components. Two trials were run using both good and poor initial values. The number of iterations for the data augmentation algorithm was set at 60. Table 5-3 presents the results of these two trials.

Table 5-3

DA Posterior Means Resulting from Good and Poor Initial Values

	Parameters			
	σ^2	τ_{00}	τ_{01}	τ_{11}
Trial A				
Good Starters (GS)	1.01478	0.34972	-0.10295	0.03916
Poor Starters (PS)	0.50739	0.69944	-0.20590	0.07832
DA Using GS	1.03645	0.32456	-0.10815	0.04721
DA Using PS	1.03683	0.35223	-0.11799	0.05420
EM	1.01109	0.35021	-0.13314	0.06238
Trial B				
Good Starters (GS)	1.03876	0.20666	0.05190	0.10172
Poor Starters (PS)	0.51938	0.41332	0.10380	0.20344
DA Using GS	1.10300	0.15031	0.03918	0.07381
DA Using PS	1.10395	0.16144	0.04336	0.08022
EM	1.03702	0.20970	0.04367	0.10950
Population Values	1.00000	0.35517	0.00000	0.03552

DA results based on 60 iterations.

Table 5-3 provides evidence that the data augmentation algorithm was robust with respect to the initial values selected. The results from good and poor initial values were very similar.

Comparing the Results of the Two Statistical Procedures

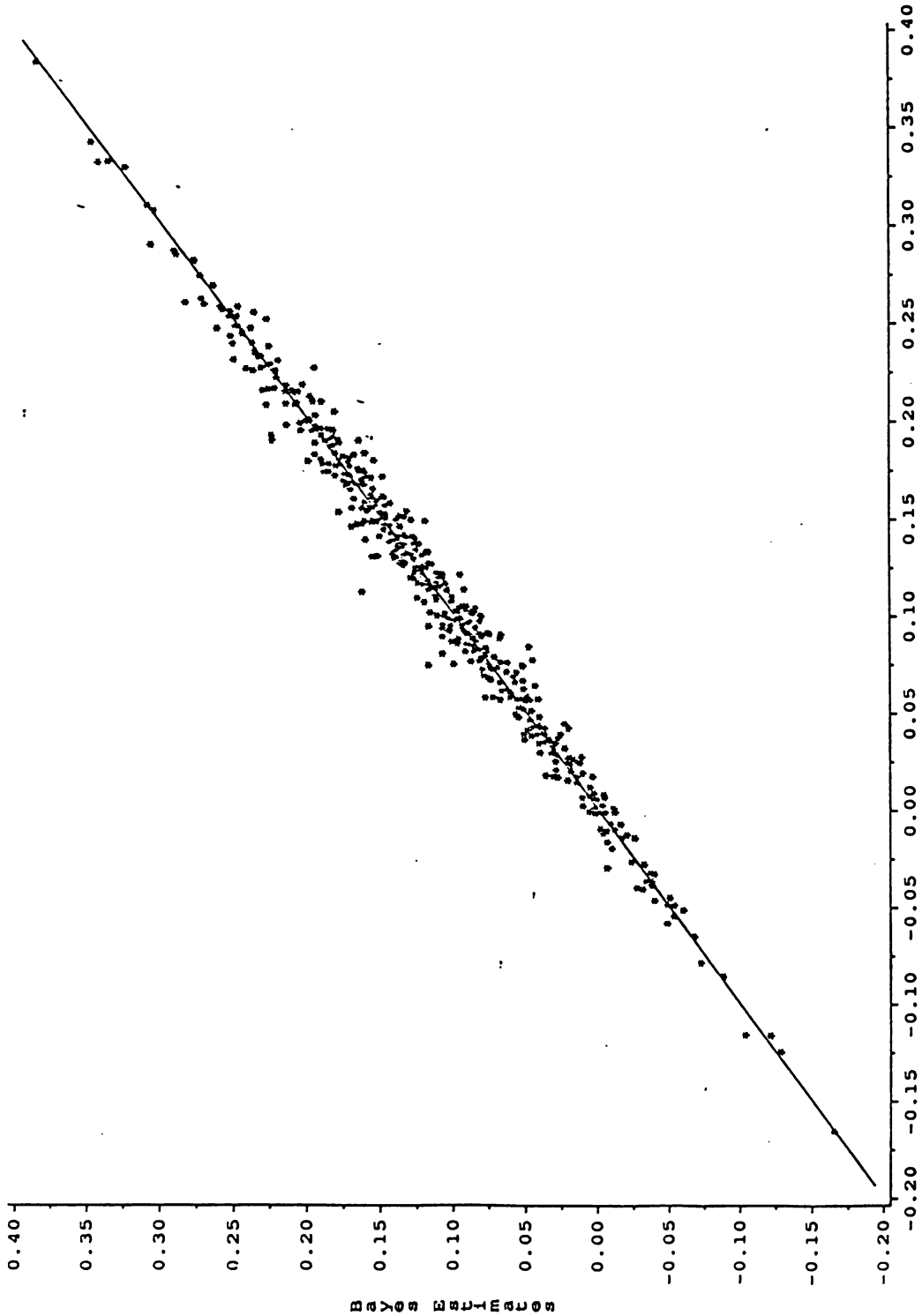
The three experimental conditions reflect changes in the interaction effect of a student level characteristic and the experimental treatment applied at the class level. The population value of γ_{11} increases from condition 1 to 3 while the conditional variance of τ_{11} decreases. The remaining parameters did not vary across experimental conditions. The results pertinent for evaluating how well these

procedures recover the parameters of interest across these three experimental conditions are summarized in this section.

Ratios of the mean squared errors were used for inferences about the relative accuracy of the two statistical procedures over the 500 experimental replications. These ratios formed F test statistics to test for the equality of the mean squared errors versus the alternative that they were not equal. These F statistics were compared to the tabled F value with 500 degrees of freedom associated with the numerator and the denominator parameters of this central F distribution.

Tables A-1, A-3 and A-5 in Appendix A display the means for γ obtained from each experimental condition and the population values for γ are provided for reference. The mean values for γ were calculated from the 500 replications of each experimental condition. The values calculated for γ from the empirical Bayes approach using the EM algorithm and the Bayesian approach using the data augmentation algorithm were very similar. The ratio of mean squared errors for γ obtained from the two procedures relative to the population parameter values are presented in tables A-7, A-8, and A-9 in Appendix A. All of the F statistics formed by these ratios were not significant at the 0.05 level. Therefore, the estimates of the γ vector obtained by the two statistical procedures were not significantly different at the 0.05 level. To illustrate how close the Bayes and empirical Bayes estimates were, Figure 5-1 displays a scatterplot of the estimates for the experimental condition 3 ATI effect parameter γ_{11} for the two procedures. The correlation between the estimates from the two procedures was 0.992.

Figure 5-1
Scatterplot of Condition 3 ATI Effect



The mean values for the variance-covariance components calculated from the 500 replications of each experimental condition are presented in Appendix A in tables A-2, A-4 and A-6. The population values are provided in these three tables for reference. The mean squared errors of both statistical procedures relative to the values of the population parameters were calculated for σ^2 and T over the 500 replications and displayed for the three experimental conditions in tables A-10, A-11 and A-12 in Appendix A. The F tests comparing the Bayesian and empirical Bayes point estimates indicated that there was a statistical difference ($p < 0.01$) between the two methods in their ability to recover the τ_{11} variance component for all three experimental conditions. The evidence suggests that the Bayesian approach using the data augmentation algorithm obtained posterior mean estimates of τ_{11} that were closer to the population value. In addition, the Bayesian approach obtained posterior mean estimates under condition 3 of the covariance term τ_{01} that were closer to the population value with a significance probability less than 0.05.

Type I Error Rate Results

The actual type I error rates were compared to the nominal rates at three levels of significance, 0.01, 0.05 and 0.10. The Bayesian approach constructs a test from the upper and lower $\alpha/2$ percentile points of the posterior distribution of γ. The empirical Bayes approach provided standard z-tests and t-tests. There were three parameters, γ_{00}, γ_{01} and γ_{11}, that had population values of zero for experimental condition 1. The parameters γ_{00} and γ_{01} have population values equal to zero for experimental conditions 2 and 3. Thus, for the three experimental conditions there were seven parameters to be tested at three nominal levels of

significance. This resulted in a total of twenty-one tests for each hypothesis testing procedure. To provide an index for ranking the relative performance, the number of times a test procedure obtained a type I error rate that was not significantly different from the nominal rate with $\alpha = 0.05$ was tabulated. Thus, an index value of 21 for a given procedure indicates that all comparisons showed no significant statistical difference between the obtained and the nominal type I error rates. This index was tabulated from tables A-13 through A-19 listed in Appendix A. In Table 5-4 the performance of the hypothesis testing methods are ranked according to this index.

Table 5-4

Performance Ranking of Hypothesis Testing Procedures

Procedure	Rank	Index
EM t-test	1	19
EM z-test	2	6
DA	3	4

According to the index rankings presented in Table 5-4, the empirical Bayes approach using a t-test to test the γ parameters that had a population value of zero provides evidence that this testing method produces an observed type I error rate that was much closer to the nominal rate than the other two testing procedures. The Bayes procedure and the empirical Bayes procedure using the z-test appear to be similar in terms of there index values listed in Table 5-4. Both of these tests were liberal and they tended to reject the null hypothesis more often than expected.

The type I error rates may also be compared with McNemar's test. A chi-square statistic is produced by McNemar's test that can be compared to the central chi-square value with one degree of freedom and $\alpha = 0.05$. The critical value for this test is 3.84.

The results from twenty-one McNemar's tests are presented in tables A-20 through A-28. The Bayesian approach was statistically different from the empirical Bayes approach using the t-test with $p < 0.05$ in 18 out of the 21 tests. The Bayesian approach was different from the empirical Bayes approach using the z-test in 6 out of the 21 tests. These six tests that indicated a statistical difference were for tests of type I error rates with the nominal error rate set at 0.10.

Power Analysis Results

There are two elements of γ that had true effects in this simulation study. The first element, γ_{10}, represented the intercept or base of the β_{j1}'s. This effect was detected with $\alpha = 0.05$ by the three testing procedures for all replications of the three experimental conditions. As expected, this effect was detected with 100% power by all three testing procedures.

The other parameter of interest in this power analysis was γ_{11}. The observed power of each test for γ_{11} with $\alpha = 0.05$ was calculated for the two experimental conditions that had true effects. The power of the empirical Bayes approach using both the z-test and the t-test were compared to the power of the Bayesian approach. The correlation between β_{j1} and W_j was set at 0.30 for experimental condition 2 and this produced a relatively small value for γ_{11}. This correlation was increased to 0.60 for experimental condition 3 and as a result, the power to detect γ_{11} was

expected to be greater. In both cases, the power of the test was expected to be relatively weak, since the proportion of correct decisions was expected to be small. The following table displays the observed power of the statistical tests for experimental conditions 2 and 3.

Table 5-5

Power of the Test of γ_{11} for Experimental Conditions 2 and 3

	Condition	
Method	2	3
DA	0.124	0.298
EM z-test	0.120	0.298
EM t-test	0.066	0.202

Power based on 500 replications.

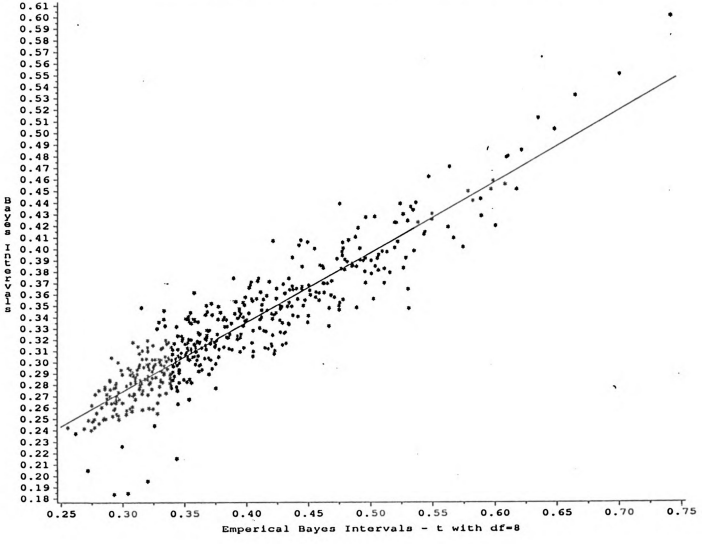
Table 5-5 illustrates the weak power these tests have for the small sample conditions of this study. This weak power was expected. In both experimental conditions, the Bayes approach using the data augmentation algorithm and the empirical Bayes approach using the EM algorithm and the z-test appeared to be slightly more powerful in detecting a true effect when compared to the empirical Bayes approach using the t-test. As noted earlier, these two more powerful testing procedures had observed type I error rates that were generally greater than the nominal rate. They appear more powerful because they were liberal tests.

From McNemar's tests presented in tables A-29 and A-30 in Appendix A, the Bayesian approach and the empirical Bayes approach using the t-test were

statistically different with significance probability $p < 0.01$ for both experimental conditions. When the Bayesian approach was compared to the empirical Bayes approach using the z-test, there was no significant difference between these procedures at $\alpha = 0.05$ for both experimental conditions.

To help illustrate the relationship between the Bayesian and empirical Bayes t-test hypothesis testing procedures, the 95% HPD intervals about the ATI effect parameter γ_{11} for experimental condition 3 are presented in Figure 5-2. The correlation between the intervals produced by these two methods was 0.923, which was based on 500 replications of this condition.

Figure 5-2
Scatterplot of 95% HPD Intervals for Condition 3 ATI Effect



Computational Efficiency Results

The amount of computer resources required by the data augmentation algorithm is an important consideration for the application of this method to substantive problems in education. The following table presents the central processing unit time on an IBM 3090-180 mainframe computer utilized by the data augmentation algorithm and the EM algorithm.

Table 5-6

Central Processing Unit Time

Experimental Conditions	Replications	CPU Seconds DA	CPU Seconds EM
1	400	9136	78
2	500	18676	95
3	500	14180	123
Totals	1400	41992	296
Average		29.99	0.21

The EM algorithm implemented in the HLM computer program (Bryk et al., 1986) has been available to researchers for three years and it has been applied to many hierarchical data sets. It can be seen clearly from Table 5-6 that the processing time required by the EM algorithm is well within the practical limits for everyday use by researchers, whereas the data augmentation algorithm requires substantial computer resources and may be rather expensive to run repeatedly. One point of interest was whether variations in the parameters affected the number of iterations that the EM algorithm requires to converge. In particular, does

decreasing the conditional value of T increase the number of iterations necessary to attain convergence? The average number of iterations required for the EM algorithm to converge for the three experimental conditions are presented in the following table.

Table 5-7

The Average Number of Iterations of the EM Algorithm

Experimental Condition	Average Iterations
1	41.346
2	40.790
3	53.678

500 replications per condition.

There appeared to be a increase in the number of iterations required for convergence for experimental condition 3. This may have been due to the fact that the conditional variance of τ_{11} was the smallest in experimental condition 3 and it is closest to the boundary point of a variance component's parameter space. This may have caused some instability estimating this variance parameter and as a consequence, the EM algorithm required more iterations to converge.

CHAPTER VI

DISCUSSION

The results presented in the previous chapter are discussed and summarized in this chapter. The following section discusses some of the advantages and disadvantages of the Bayesian and empirical Bayes implementations. And finally, some unanswered questions and suggestions for future research are presented.

Advantages and Disadvantages

The fundamental question of this simulation study was whether or not the Bayesian approach implemented with the data augmentation algorithm would recover the parameters of the hierarchical model more accurately than the empirical Bayes approach using the EM algorithm. The performance of these two statistical methodologies was primarily evaluated in terms of their mean squared errors relative to the population parameter values. In all experimental conditions investigated there was no statistical difference between point estimates produced by the Bayesian procedure using the data augmentation algorithm and the empirical Bayes procedure using the EM algorithm in terms of recovering γ and σ^2. The Bayesian approach did produce posterior mean estimates for some elements of T that were substantially closer to their population values than the empirical Bayes approach. The T matrix was estimated from a small sample and the Bayesian approach was superior in

recovering this variance-covariance matrix. Therefore, we may conclude that the Bayes approach implemented with the data augmentation algorithm recovered the parameters of the model under investigation as well or better than the empirical Bayes approach using the EM algorithm.

Another advantage of the Bayesian approach implemented with the data augmentation algorithm is that it produces a finite approximation to the posterior distribution. The information supplied by the complete posterior distribution provides the researcher with a number of options. A variety of measures from a posterior distribution may be easily determined, for example, the mean, median, mode and percentile points. In addition, it is possible to graph an entire posterior distribution of a parameter.

The empirical Bayes procedure using a t-test was clearly superior for hypothesis tests of the elements of γ that had population values of zero. The observed type I error rates from this procedure were very close to the nominal rates. The Bayesian procedure as implemented in this study and the empirical Bayes approach using the z-test were liberal and tended to reject the null hypothesis more often then the expected nominal rate. The Bayesian procedure using the data augmentation algorithm may produce a type I error rate closer to the nominal rate if the maximum number of augmented data sets (m) was increased. The tail sections of a posterior distribution might be approximated more accurately with a larger number of augmented data sets.

The power to detect γ_{11}, the interaction effect between the student level independent measure and the between-class treatment, was weak for both experimental conditions that had a true effect associated with this parameter. Since

there were only ten classes, the power of this test was expected to be weak. The observed power for the Bayes and empirical Bayes procedure using the z-test were approximately equivalent and they were more powerful than the empirical Bayes procedure using a t-test. The two procedures that appeared more powerful achieved this power at the expense of controlling type I errors. Their type I error rates were higher than the expected nominal rate.

The primary disadvantage of the data augmentation algorithm was the large amount of computer resources it requires. Each replication took an average of about 30 seconds to analyze on an IBM 3090-180 mainframe computer. This computer is a relatively fast general purpose mainframe computer. The empirical Bayes approach using the EM algorithm was much more efficient in terms of its computer resource requirements. The HLM computer program (Bryk et al., 1986) required on the average only 0.21 seconds to converge.

One unresolved problem with the data augmentation algorithm is that there is no general strategy for determining convergence. This makes it less appealing to researchers working on substantive problems in education. Also, the implementation of the data augmentation algorithm created for this study was very specific in nature. It was designed for the one model studied in this simulation and it is not generally available for use by researchers in education.

Suggestions for Future Research

This simulation study attempted to model conditions that might be obtained in an education research project focused on student by class level treatment interaction effects. Consequently, the parameters for the experimental data sets were

selected to reflect typical data that might be observed. The class size was selected to model the class size distribution of elementary schools. This distribution of class size was not that variable and as a result the overall sampling design was only moderately unbalanced. It is plausible that as the sampling distribution of class size becomes more variable, the results between the two statistical procedures will increasingly diverge. The distribution of class size was not a variable in this study. Investigation of the influence different class size distributions have may provide insight into situations that capitalize on the strengths of the Bayesian method. Since the data augmentation algorithm provides a new tool for a true Bayesian approach to data analysis, additional experience in a wide variety of situations is important to fully evaluate its usefulness. It would be extremely informative to know under what general conditions the data augmentation algorithm is clearly superior and under what conditions other statistical procedures are adequate. This is a concern given the large amount of computer resources required by the data augmentation algorithm.

There are a number of different ways that the data augmentation algorithm can be implemented. The number of augmented data sets and how they are increased is a variable of the algorithm that may effect the rate of convergence. The number of augmented data sets (m) was doubled until it reached a maximum of 2048. The number of data sets and the rate at which they are increased could be varied and investigated to determine an optimal strategy for controlling the augmentation of data sets. In addition, if the maximum number of augmented data sets is increased, will the observed type I error rate become closer to the nominal error rate? This algorithm may require a large number of augmented data sets to

accurately reflect the tail areas of a posterior distribution.

Developing a general stopping rule for the data augmentation algorithm is an important and challenging issue for future investigation. Since this was a simulation study and the population parameters were known, a stopping rule was developed empirically. This approach is not much help in applied situations where the parameter values are unknown and the conditions of the data analysis project are not similar to the conditions investigated in this simulation study. A stopping criterion must consider convergence approaching some finite stochastic limit within some tolerance and at the same time investigate trends across iterations. This is a much more difficult task than traditional stopping rules that examine some function of differences between iterations for a value less than some predetermined criterion.

The cost for computer resources required by the data augmentation algorithm will decrease in the future because computers are getting faster and more inexpensive. In addition, the specific implementation of the data augmentation procedure could be optimized to run faster by rewriting sections of the computer code. Since this was the first implementation of this method for the hierarchical model under investigation, many parts of the code were developed to clearly and accurately reflect the algorithm, and optimization was a secondary concern. Future simulation studies investigating properties of the data augmentation algorithm could be implemented on a parallel processing computer. Simulation studies like this one are suitable for parallel processing computers because every replication is independent and may be processed separately.

For the educational researcher considering methods to analyze experimental data that are expected to be similar to the data investigated in this simulation study,

the empirical Bayes procedure implemented with the EM algorithm and using a t-test for the γ effects is recommended at this time. This method is recommended because there are computer programs available that can analyze a large variety of models, the cost for the required computer resources is substantially less, and it appears to perform well for hypotheses tests. Possibly, for other analysis situations the Bayesian procedure implemented with the data augmentation algorithm will prove to be clearly superior. It would be interesting to evaluate the performance of the Bayesian approach using the data augmentation algorithm in a situation with larger variability in group (class) size. In the past, the Bayesian approach has been held back because there were no practical means for its implementation. Now with the data augmentation algorithm as a practical means for implementing this approach researchers can begin to acquire some experience. And with this experience, researchers will become more aware of the advantages and disadvantages of these two statistical approaches.

APPENDICES

APPENDIX A

TABLES CONTAINING SIMULATION RESULTS

Table A-1

Posterior Estimates for γ from Experimental Condition 1

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA Mean	0.00992	0.00183	0.75767	0.00401
EM Mean	0.00984	0.00201	0.75713	0.00389
DA Sd of Mean	0.19627	0.20402	0.08926	0.08639
EM Sd of Mean	0.19609	0.20366	0.08974	0.08670
Corr between DA and EM	0.99948	0.99951	0.99474	0.99332
Population Values	0.00000	0.00000	0.76320	0.00000

Estimates based on 500 replications.

Table A-2

Posterior Estimates for σ^2 and T from Experimental Condition 1

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	1.02767	0.34374	0.00060	0.03171
EM	0.99654	0.36984	0.00000	0.04172
Population Values	1.00000	0.35517	0.00000	0.03552

Estimates based on 500 replications.

Table A-3

Posterior Estimates for γ from Experimental Condition 2

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA Mean	-0.00160	-0.00653	0.75381	0.06044
EM Mean	-0.00142	-0.00646	0.75422	0.06007
DA Sd of Mean	0.19503	0.21104	0.08737	0.08861
EM Sd of Mean	0.19482	0.21058	0.08695	0.08847
Corr between DA and EM	0.99948	0.99952	0.99309	0.99124
Population Values	0.00000	0.00000	0.76320	0.05654

Estimates based on 500 replications.

Table A-4

Posterior Estimates for σ^2 and T from Experimental Condition 2

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	1.02972	0.32566	-0.00271	0.02828
EM	0.99875	0.35238	-0.00248	0.03731
Population Values	1.00000	0.35517	0.00000	0.03232

Estimates based on 500 replications.

Table A-5

Posterior Estimates for γ from Experimental Condition 3

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA Mean	-0.00969	0.00138	0.75950	0.11602
EM Mean	-0.00945	0.00111	0.75990	0.11616
DA Sd of Mean	0.20464	0.19215	0.08609	0.08335
EM Sd of Mean	0.20452	0.19232	0.08463	0.08256
Corr between DA and EM	0.99960	0.99957	0.99116	0.99199
Population Values	0.00000	0.00000	0.76320	0.11308

Estimates based on 500 replications.

Table A-6

Posterior Estimates for σ^2 and T from Experimental Condition 3

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	1.01760	0.34005	0.00064	0.02353
EM	0.98989	0.36513	0.00041	0.02971
Population Values	1.00000	0.35517	0.00000	0.02273

Estimates based on 500 replications.

Table A-7

MSE's Relative to Population Values for γ from Condition 1

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA	0.03854299	0.04154461	0.00798253	0.00746498
EM	0.03847160	0.04139983	0.00807444	0.00751721
Ratio DA/EM	1.0019	1.0035	0.9886	0.9931

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-8

MSE's Relative to Population Values for γ from Condition 2

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA	0.03796498	0.04449211	0.00770650	0.00785097
EM	0.03788249	0.04429731	0.00762598	0.00782378
Ratio DA/EM	1.0022	1.0044	1.0106	1.0035

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-9

MSE's Relative to Population Values for γ from Condition 3

	Parameters			
Method	γ_{00}	γ_{01}	γ_{10}	γ_{11}
DA	0.04188586	0.03685111	0.00740986	0.00694154
EM	0.04183597	0.03691413	0.00715952	0.00681267
Ratio DA/EM	1.0012	0.9983	1.0350	1.0189

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-10

MSE's Relative to Population Values for σ^2 and T from Condition 1

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	0.01048149	0.04103996	0.00361215	0.00077421
EM	0.00929093	0.04051690	0.00424074	0.00147059
Ratio DA/EM	1.1281	1.0129	0.8518	0.5265 **

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-11

MSE's Relative to Population Values for σ^2 and T from Condition 2

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	0.01034900	0.03917907	0.00304788	0.00061088
EM	0.00885991	0.03749094	0.00350264	0.00129564
Ratio DA/EM	1.1691	1.0450	0.8702	0.4715 **

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-12

MSE's Relative to Population Values for σ^2 and T from Condition 3

	Parameters			
Method	σ^2	τ_{00}	τ_{01}	τ_{11}
DA	0.00977809	0.03958772	0.00257984	0.00046431
EM	0.00881106	0.03869871	0.00314165	0.00106294
Ratio DA/EM	1.1098	1.0230	0.8212 *	0.4368 **

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-13

Type I Error Rates for γ_{00} from Experimental Condition 1

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.018	1.798	0.096	4.720 **	0.152	3.876 **
EM z-test	0.020	2.247 *	0.086	3.694 **	0.134	2.534 *
EM t-test	0.006	-0.899	0.040	-1.026	0.096	-0.298

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-14

Type I Error Rates for γ_{01} from Experimental Condition 1

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.032	4.944 **	0.082	3.283 **	0.140	2.981 **
EM z-test	0.028	4.045 **	0.080	3.078 **	0.122	1.640
EM t-test	0.014	0.899	0.046	-0.410	0.090	-0.745

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-15

Type I Error Rates for γ_{11} from Experimental Condition 1

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.008	-0.450	0.046	-0.410	0.098	-0.149
EM z-test	0.012	0.450	0.054	0.410	0.104	0.298
EM t-test	0.004	-1.348	0.022	-2.873 **	0.072	-2.087 *

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-16

Type I Error Rates for γ_{00} from Experimental Condition 2

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.034	5.394 **	0.088	3.899 **	0.138	2.832 **
EM z-test	0.032	4.944 **	0.076	2.668 **	0.120	1.491
EM t-test	0.006	-0.899	0.048	-0.205	0.100	0.000

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-17

Type I Error Rates for γ_{01} from Experimental Condition 2

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.036	5.843 **	0.112	6.361 **	0.180	5.963 **
EM z-test	0.038	6.293 **	0.110	6.156 **	0.162	4.621 **
EM t-test	0.014	0.899	0.060	1.026	0.124	1.789

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-18

Type I Error Rates for γ_{00} from Experimental Condition 3

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.028	4.045 **	0.086	3.694 **	0.144	3.280 **
EM z-test	0.034	5.394 **	0.086	3.694 **	0.128	2.087 *
EM t-test	0.018	1.798	0.036	-1.436	0.098	-0.149

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-19

Type I Error Rates for γ_{01} from Experimental Condition 3

	$\alpha = 0.01$		$\alpha = 0.05$		$\alpha = 0.10$	
Method	Rate	z	Rate	z	Rate	z
DA	0.022	2.697 **	0.084	3.488 **	0.140	2.981 **
EM z-test	0.026	3.596 **	0.080	3.078 **	0.118	1.342
EM t-test	0.004	-1.348	0.038	-1.231	0.088	-0.894

Note. * $p < 0.05$; ** $p < 0.01$ (two-tailed test significance levels). Based on 500 replications.

Table A-20

McNemar's Statistics for Condition 1 Type I Errors with Nominal $\alpha = 0.01$

	Parameters		
DA vs. Method	γ_{00}	γ_{01}	γ_{11}
EM z-test	0.000	0.500	0.500
EM t-test	4.167 *	7.111 **	0.500

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-21

McNemar's Statistics for Condition 1 Type I Errors with Nominal $\alpha = 0.05$

DA vs. Method	Parameters		
	γ_{00}	γ_{01}	γ_{11}
EM z-test	3.200	0.000	0.643
EM t-test	26.036 **	16.056 **	8.643 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-22

McNemar's Statistics for Condition 1 Type I Errors with Nominal $\alpha = 0.10$

DA vs. Method	Parameters		
	γ_{00}	γ_{01}	γ_{11}
EM z-test	5.818 *	7.111 **	0.308
EM t-test	26.056 **	23.040 **	8.471

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-23

McNemar's Statistics for Condition 2 Type I Errors with Nominal $\alpha = 0.01$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	0.000	0.000
EM t-test	12.071 **	9.091 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-24

McNemar's Statistics for Condition 2 Type I Errors with Nominal $\alpha = 0.05$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	3.125	0.000
EM t-test	18.050 **	24.039 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-25

McNemar's Statistics for Condition 2 Type I Errors with Nominal $\alpha = 0.10$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	7.111 **	5.818 *
EM t-test	17.053 **	26.036 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-26

McNemar's Statistics for Condition 3 Type I Errors with Nominal $\alpha = 0.01$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	1.333	0.167
EM t-test	3.200	7.111 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-27

McNemar's Statistics for Condition 3 Type I Errors with Nominal $\alpha = 0.05$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	0.250	0.500
EM t-test	23.040 **	21.044 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-28

McNemar's Statistics for Condition 3 Type I Errors with Nominal $\alpha = 0.10$

DA vs. Method	Parameters	
	γ_{00}	γ_{01}
EM z-test	6.125 *	7.692 **
EM t-test	21.044 **	24.039 **

Note. * $p < 0.05$; ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-29

Experimental Condition 2 Power Comparisons with DA for γ_{11}

Method	χ^2
EM z-test	0.071
EM t-test	25.290 **

Note. ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

Table A-30

Experimental Condition 3 Power Comparisons with DA for γ_{11}

Method	χ^2
EM z-test	0.033
EM t-test	46.021 **

Note. ** $p < 0.01$ for one degree of freedom test. Based on 500 replications.

APPENDIX B

SOURCE CODE FOR COMPUTER PROGRAMS

Data Augmentation Algorithm FORTRAN Source Code

```
      PROGRAM DAUG

************************************************************************
*     This program calculates the posterior distribution of the
*     between-group regression coefficients for a two level
*     hierarchical model using the data augmentation algorithm.

      PARAMETER  (MXAUG=2048, MXIT=25, MXK=10)
      INTEGER  ISEED, IT, MPT, NEXP, NGEN, NK, NT, NTOT
      LOGICAL  DONE
      REAL  B(2,MXK,MXAUG), BLS(2,MXK), DWWIN(4)
      REAL  G(4,MXAUG), GAMMA(4), ID2(2,2), P0W, P1W, SIG(MXAUG), SIGMA2
      REAL  SXY(2,MXK), SY2(MXK), TAU(3), TIN(2,2,MXAUG), TZ(3), WC(2,4)
      REAL  WE(2,4), WI(MXK), XX(2,2,MXK), XXIN(2,2,MXK)
      COMMON  /BCOM/ B /GCOM/ G /SIGCOM/ SIG /TINCOM/ TIN

      OPEN(10, FILE='SSM', FORM='UNFORMATTED')
      OPEN(20, FILE='INIT')
      OPEN(25, FILE='VARS')
      OPEN(30, FILE='OUT1')
      OPEN(40, FILE='LAST')
      OPEN(50, FILE='SEED')
      OPEN(60, FILE='KNOW')

*     Prepare IMSL random number generators.
      CALL RNOPT(3)
      ISEED = 123457
      CALL RNSET(ISEED)
      NTOT = 0
      NK = 10

      READ(20,201) NEXP, P0W, P1W
      CALL DESIGN(DWWIN, ID2, NK, WC, WE)

      DO 20 N = 1, NEXP
        CALL READER(BLS, NK, NT, SXY, SY2, TZ, WI, XX, XXIN)
        CALL KNOWN(BLS, GAMMA, P0W, P1W, SIGMA2, TAU, WI, XXIN)
        IT = 1
        MPT = 1
        DONE = .FALSE.
*       Begin DO-WHILE loop.
111     CONTINUE
```

```
        IF (DONE) GOTO 888
          CALL GSTAR(DWWIN, IT, MPT, NK, WC, WE, WI)
          IF (2 * MPT .LE. MXAUG) THEN
            NGEN = 2
            CALL SIMG(IT, MPT, NGEN, NK)
            MPT = 2 * MPT
          ELSE
            NGEN = 1
            CALL SIMG(IT, MPT, NGEN, NK)
          END IF
          CALL TSTAR(IT, MPT, NGEN, NK, WC, WE, WI)
          CALL SIMT(IT, MPT, NK, TZ)
          CALL BETA(BLS, ID2, MPT, NGEN, NK, WC, WE, WI, XX)
          CALL SSTAR(IT, MPT, NK, NT, SXY, SY2, XX)
          CALL SIMS(IT, MPT, NT)

*         Update loop control.
          IF (IT .EQ. MXIT) THEN
            DONE = .TRUE.
          ELSE
            IT = IT + 1
          END IF

        GOTO 111
888     CONTINUE
*       End DO-WHILE loop.

*     Sum the total number of experimental units for all experiments.
      NTOT = NTOT + NT
 20   CONTINUE

      CALL NTAVE(NEXP, NTOT)
      CALL RNGET(ISEED)
      WRITE(50,501) ISEED

201   FORMAT(I5,2F5.2)
501   FORMAT(I10)

      END
```

```
      SUBROUTINE DESIGN(DWWIN, ID2, NK, WC, WE)
*************************************************************************
*     Construct ID2 and the diagonal matrix W'W inverse.  Also,
*     construct the two between-group design matrices.  WC indicates
*     the control and WE indicates the experimental conditions.

      INTEGER  NK
      REAL  DWWIN(4), ID2(2,2), WC(2,4), WE(2,4)

      DO 10 I = 1, 4
        DWWIN(I) = 1.0 / NK
 10   CONTINUE

      ID2(1,1) = 1.0
      ID2(2,1) = 0.0
      ID2(1,2) = 0.0
      ID2(2,2) = 1.0

      WC(1,1) =  1.0
      WC(2,1) =  0.0
      WC(1,2) = -1.0
      WC(2,2) =  0.0
      WC(1,3) =  0.0
      WC(2,3) =  1.0
      WC(1,4) =  0.0
      WC(2,4) = -1.0

      WE(1,1) = 1.0
      WE(2,1) = 0.0
      WE(1,2) = 1.0
      WE(2,2) = 0.0
      WE(1,3) = 0.0
      WE(2,3) = 1.0
      WE(1,4) = 0.0
      WE(2,4) = 1.0

      RETURN
      END
```

```
      SUBROUTINE READER(BLS, NK, NT, SXY, SY2, TZ, WI, XX, XXIN)
**********************************************************************
*     Read sufficient sum of squares statistics from logical unit 10
*     and the initial estimates for Beta, Sigma_2, and Tau from logical
*     unit 20.  Calculate Y'Y and Inv(Tau).  Construct for each group
*     X'Y, Inv(X'X) and XX.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  NJ(MXK), NK, NT
      REAL  B(2,MXK,MXAUG), BLS(2,MXK), DET, FS, FT(2,2), SIG(MXAUG)
      REAL  SX(MXK), SXY(2,MXK), SX2(MXK), SY(MXK), SY2(MXK), T(2,2)
      REAL  TIN(2,2,MXAUG), TZ(3), WI(MXK), XX(2,2,MXK), XXIN(2,2,MXK)
      DOUBLE PRECISION  DSX, DSXY, DSX2, DSY, DSY2, DWI
      CHARACTER *12  ID
      COMMON  /BCOM/ B /SIGCOM/ SIG /TINCOM/ TIN

*     SSM data file must be converted from DOUBLE to SINGLE PRECISION.

      DO 10 K = 1, NK
        READ(20,201) (BLS(I,K), I = 1, 2)
 10   CONTINUE
      READ(20,202) SIG(1)
      DO 20 I = 1, 2
        READ(20,203) (T(I,J), J = 1, 2)
 20   CONTINUE
      TZ(1) = T(1,1)
      TZ(2) = T(2,1)
      TZ(3) = T(2,2)

*     Reading final EM estimates for Sigma_2 and Tau.
      READ(20,204) FS
      DO 30 I = 1, 2
        READ(20,203) (FT(I,J), J = 1, 2)
 30   CONTINUE
      READ(20,205)

*     Calculate TIN = Inv(Tau).
      DET = T(1,1) * T(2,2) - T(2,1) * T(1,2)
      TIN(1,1,1) = T(2,2) / DET
      TIN(2,1,1) = -1.0 * T(2,1) / DET
      TIN(1,2,1) = TIN(2,1,1)
      TIN(2,2,1) = T(1,1) / DET

      DO 40 M = 1, 3
        READ(10)
 40   CONTINUE
      NT = 0.0
      DO 50 K = 1, NK
        READ(10) NJ(K)
        NT = NT + NJ(K)
        READ(10) DSY, DSX
        SY(K) = SNGL(DSY) * NJ(K)
        SX(K) = SNGL(DSX) * NJ(K)
        SXY(1,K) = SY(K)
        READ(10) DSY2, DSXY, DSX2
        DSY2 = DSY2 + NJ(K) * DSY * DSY
        DSXY = DSXY + NJ(K) * DSY * DSX
        DSX2 = DSX2 + NJ(K) * DSX * DSX
```

```
        SY2(K) = SNGL(DSY2)
        SXY(2,K) = SNGL(DSXY)
        SX2(K) = SNGL(DSX2)
        READ(10) ID, DWI
        WI(K) = SNGL(DWI)

*       Construct (X'X) for each group.
        XX(1,1,K) = NJ(K)
        XX(2,1,K) = SX(K)
        XX(1,2,K) = SX(K)
        XX(2,2,K) = SX2(K)

*       Calculate Inv(X'X)
        DET = XX(1,1,K) * XX(2,2,K) - XX(2,1,K) * XX(1,2,K)
        XXIN(1,1,K) = XX(2,2,K) / DET
        XXIN(2,1,K) = -1.0 * XX(2,1,K) / DET
        XXIN(1,2,K) = XXIN(2,1,K)
        XXIN(2,2,K) = XX(1,1,K) / DET

*       Initialize B in BCOM common block.
        DO 60 I = 1, 2
          B(I,K,1) = BLS(I,K)
 60     CONTINUE
 50   CONTINUE

201   FORMAT(30X,2F12.5)
202   FORMAT(5(/),16X,E13.6/)
203   FORMAT(8X,2F12.5)
204   FORMAT(/16X,E13.6/)
205   FORMAT(7(/))

      RETURN
      END
```

```
      SUBROUTINE KNOWN(BLS, GAMMA, P0W, P1W, SIGMA2, TAU, WI, XXIN)
*************************************************************************
*     Calculate Gamma and D(Gamma) from the knowledge of the
*     complete data.

      PARAMETER(MXK=10)
      REAL  BLS(2,MXK), D(2,2), DINB(2), DET, DIN(2,2), DIS(4,4)
      REAL  GAMMA(4), P0W, P1W, SIGMA2, SWDINB(4), T(2,2), TAU(3)
      REAL  WI(MXK), XXIN(2,2,MXK)

      READ(25,201) SIGMA2
      READ(25,202) (TAU(J), J = 1, 3)
      T(1,1) = TAU(1)
      T(2,1) = TAU(2)
      T(1,2) = TAU(2)
      T(2,2) = TAU(3)
      DO 10 J = 1, 4
        SWDINB(J) = 0.0
        DO 10 I = 1, 4
          DIS(I,J) = 0.0
 10   CONTINUE

      DO 20 K = 1, MXK
*       Calculate DIN = Inv(D) = Inv(T + V) for each group.
        DO 30 J = 1, 2
          DO 30 I = 1, 2
            D(I,J) = T(I,J) + XXIN(I,J,K) * SIGMA2
 30     CONTINUE
        DET = D(1,1) * D(2,2) - D(1,2) * D(2,1)
        DIN(1,1) = D(2,2) / DET
        DIN(2,1) = -1.0 * D(2,1) / DET
        DIN(1,2) = DIN(2,1)
        DIN(2,2) = D(1,1) / DET

*       Calculate Sum(DIS) = Sum(W'Inv(D)W) across groups.
        DIS(1,1) = DIS(1,1) + DIN(1,1)
        DIS(2,1) = DIS(2,1) + WI(K) * DIN(1,1)
        DIS(3,1) = DIS(3,1) + DIN(2,1)
        DIS(4,1) = DIS(4,1) + WI(K) * DIN(2,1)
        DIS(1,2) = DIS(1,2) + WI(K) * DIN(1,1)
        DIS(2,2) = DIS(2,2) + DIN(1,1)
        DIS(3,2) = DIS(3,2) + WI(K) * DIN(2,1)
        DIS(4,2) = DIS(4,2) + DIN(2,1)
        DIS(1,3) = DIS(1,3) + DIN(1,2)
        DIS(2,3) = DIS(2,3) + WI(K) * DIN(1,2)
        DIS(3,3) = DIS(3,3) + DIN(2,2)
        DIS(4,3) = DIS(4,3) + WI(K) * DIN(2,2)
        DIS(1,4) = DIS(1,4) + WI(K) * DIN(1,2)
        DIS(2,4) = DIS(2,4) + DIN(1,2)
        DIS(3,4) = DIS(3,4) + WI(K) * DIN(2,2)
        DIS(4,4) = DIS(4,4) + DIN(2,2)
*       Calculate SWDINB = Sum(W'Inv(D)BLS)
        DO 40 I = 1, 2
          DINB(I) = 0.0
          DO 50 J = 1, 2
            DINB(I) = DINB(I) + DIN(I,J) * BLS(J,K)
 50       CONTINUE
 40     CONTINUE
```

```
        SWDINB(1) = SWDINB(1) + DINB(1)
        SWDINB(2) = SWDINB(2) + WI(K) * DINB(1)
        SWDINB(3) = SWDINB(3) + DINB(2)
        SWDINB(4) = SWDINB(4) + WI(K) * DINB(2)
20    CONTINUE

*     Calculate DIS = Inv(Sum(W'Inv(D)W))
      CALL LINRG(4, DIS, 4, DIS, 4)
      DO 60 I = 1, 4
        GAMMA(I) = 0.0
        DO 70 J = 1, 4
          GAMMA(I) = GAMMA(I) + DIS(I,J) * SWDINB(J)
 70     CONTINUE
 60   CONTINUE

      WRITE(60,601) (GAMMA(I), I = 1, 4)
      DO 80 I = 1, 4
        WRITE(60,601) (DIS(I,J), J = 1, 4)
80    CONTINUE

201   FORMAT(F11.5)
202   FORMAT(3(F11.5))
601   FORMAT(4F11.5)

      RETURN
      END
```

```
      SUBROUTINE GSTAR(DWWIN, IT, MPT, NK, WC, WE, WI)
*************************************************************************
*     Calculating MPT vectors of Gamma_Star, where MPT is the pointer
*     that indicates the number of Beta vectors in the current
*     augmented data set.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  IT, MPT, NK, UN
      REAL  B(2,MXK,MXAUG), DWWIN(4), G(4,MXAUG)
      REAL  SWB(4), WC(2,4), WE(2,4), WI(MXK)
      COMMON /BCOM/ B /GCOM/ G

      DO 10 N = 1, MPT
        DO 20 I = 1, 4
          SWB(I) = 0.0
 20     CONTINUE
        DO 30 K = 1, NK
          IF (WI(K) .GT. 0.0) THEN
            DO 40 J = 1, 4
              DO 50 I = 1, 2
                SWB(J) = SWB(J) + WE(I,J) * B(I,K,N)
 50           CONTINUE
 40         CONTINUE
          ELSE
            DO 60 J = 1, 4
              DO 70 I = 1, 2
                SWB(J) = SWB(J) + WC(I,J) * B(I,K,N)
 70           CONTINUE
 60         CONTINUE
          END IF
 30     CONTINUE

*       Calculate the Nth Gamma vector and store it in the GCOM block.
        DO 80 I = 1, 4
          G(I,N) = DWWIN(I) * SWB(I)
 80     CONTINUE
 10   CONTINUE

301   FORMAT(4(F11.5))

      RETURN
      END
```

```
      SUBROUTINE SIMG(IT, MPT, NGEN, NK)
*************************************************************************
*     Generate random Gamma vectors.  If the current number of Gamma
*     vectors is less than MXAUG then the number of Gamma vectors will
*     be increased by 2.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  GPT, IRANK, IT, LOOPS, MPT, NGEN, NK, NR, RPT, UN
      REAL  DET, G(4,MXAUG), GBAR(4), GVAR(4), GT(4,MXAUG), R(4*MXAUG)
      REAL  S(3), SIN(3), TIN(2,2,MXAUG), U(3)
      COMMON /GCOM/ G /TINCOM/ TIN

*     Get the required i.i.d. standard normals.
      NR = 4 * NGEN * MPT
      CALL RNNOR(NR, R)
      GPT = 1
      RPT = 1

      DO 10 N = 1, MPT
*       Calculate Sum(WTINW) = S for the three values other than zero.
        S(1) = NK * TIN(1,1,N)
        S(2) = NK * TIN(1,2,N)
        S(3) = NK * TIN(2,2,N)

*       Calculate Inverse(S) = SIN
        DET = S(1) * S(3) - S(2) * S(2)
        SIN(1) = S(3) / DET
        SIN(2) = -1.0 * S(2) / DET
        SIN(3) = S(1) / DET

*       Calculate the Cholesky factor.
        U(1) = SQRT(SIN(1))
        U(2) = SIN(2) / U(1)
        U(3) = SQRT(SIN(3) - U(2) * U(2))

*       Generate Gamma from Gamma_Star, the mean of the distribution.
        DO 20 I = 1, NGEN
          GT(1,GPT) = G(1,N) + U(1) * R(RPT)
          GT(2,GPT) = G(2,N) + U(1) * R(RPT+1)
          GT(3,GPT) = G(3,N) + U(2) * R(RPT) + U(3) * R(RPT+2)
          GT(4,GPT) = G(4,N) + U(2) * R(RPT+1) + U(3) * R(RPT+3)
          RPT = RPT + 4
          GPT = GPT + 1
 20     CONTINUE
 10   CONTINUE
      LAST = NGEN * MPT
      DO 30 I = 1, 4
        DO 30 N = 1, LAST
          G(I,N) = GT(I,N)
 30   CONTINUE

*     Calculate the mean and dispersion of Gamma.
      IF (IT .EQ. 25) THEN
        UN = 30
        DO 90 I = 1, 4
          GBAR(I) = 0.0
          GVAR(I) = 0.0
 90     CONTINUE
```

```
      DO 100 I = 1, 4
        DO 100 N = 1, MXAUG
          GBAR(I) = GBAR(I) + G(I,N)
          GVAR(I) = GVAR(I) + G(I,N) * G(I,N)
100     CONTINUE
        DO 110 I = 1, 4
          GBAR(I) = GBAR(I) / REAL(MXAUG)
          GVAR(I) = GVAR(I) / REAL(MXAUG) - GBAR(I) * GBAR(I)
110     CONTINUE
        WRITE(UN,301) (GBAR(I), I = 1, 4)
        WRITE(UN,301) (GVAR(I), I = 1, 4)
        CALL TAILS(IT, UN)
      END IF

301   FORMAT(4(F11.5))

      RETURN
      END
```

```
      SUBROUTINE TSTAR(IT, MPT, NGEN, NK, WC, WE, WI)
*************************************************************************
*     TSTR = Sum(Tau_Star) / NK for each gamma vector in the augmented
*     data set.  Store these in the TINCOM block.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  IT, MPT, NGEN, NK, PPT, UN
      REAL  B(2,MXK,MXAUG), G(4,MXAUG), ST(2), TBAR(3), TIN(2,2,MXAUG)
      REAL  TSTR(2,2), WC(2,4), WCG(2), WE(2,4), WEG(2), WI(MXK)
      COMMON  /BCOM/ B /GCOM/ G /TINCOM/ TIN

      PPT = 1
      DO 10 N = 1, MPT
        DO 20 J = 1, 2
          DO 30 I = 1, 2
            TSTR(I,J) = 0.0
 30       CONTINUE
 20     CONTINUE
        DO 40 I = 1, 2
          WCG(I) = 0.0
          WEG(I) = 0.0
          DO 50 J = 1, 4
            WCG(I) = WCG(I) + WC(I,J) * G(J,N)
            WEG(I) = WEG(I) + WE(I,J) * G(J,N)
 50       CONTINUE
 40     CONTINUE

*       Calculate the vector sum of (B - WG) over groups.
        DO 60 K = 1, NK
          IF(WI(K) .LT. 0.0) THEN
            DO 70 I = 1, 2
              ST(I) = B(I,K,PPT) - WCG(I)
 70         CONTINUE
          ELSE
            DO 80 I = 1, 2
              ST(I) = B(I,K,PPT) - WEG(I)
 80         CONTINUE
          END IF
*         Let TSTR = Sum(B - WG)(B - WG)' over groups.
          DO 90 J = 1, 2
            DO 100 I = 1, 2
              TSTR(I,J) = TSTR(I,J) + ST(I) * ST(J)
100         CONTINUE
 90       CONTINUE
 60     CONTINUE

*       Store TSTR / NK in TIN
        DO 110 J = 1, 2
          DO 110 I = 1, 2
            TIN(I,J,N) = TSTR(I,J) / 10.0
110     CONTINUE
        IF (NGEN .EQ. 2 .AND. MOD(N,2) .EQ. 0) THEN
          PPT = PPT + 1
        ELSE IF (NGEN .EQ. 1) THEN
          PPT = PPT + 1
        END IF
 10   CONTINUE
*     Calculate point estimate of TSTAR and write it out.
```

```
      IF (IT .EQ. 25) THEN
        UN = 30
        DO 120 I = 1, 3
          TBAR(I) = 0.0
120     CONTINUE
        DO 130 N = 1, MXAUG
          TBAR(1) = TBAR(1) + TIN(1,1,N)
          TBAR(2) = TBAR(2) + TIN(2,1,N)
          TBAR(3) = TBAR(3) + TIN(2,2,N)
130     CONTINUE
        DO 140 I = 1, 3
          TBAR(I) = TBAR(I) / REAL(MXAUG)
140     CONTINUE
        WRITE(UN,301) (TBAR(J), J = 1, 3)
      END IF

301   FORMAT(3(F11.5))

      RETURN
      END
```

```
      SUBROUTINE SIMT(IT, MPT, NK, TZ)
*************************************************************************
*     The Tau_Star matrices currently in TINCOM are factored and
*     are used to generate new Tau inverse matrices.  These new
*     matrices are stored in the TINCOM.

      PARAMETER  (MXAUG=2048)
      INTEGER  IT, MPT, NK, UN
      LOGICAL  AGAIN
      REAL  B(3), CHIA(MXAUG), CHIB(MXAUG), CHI1, CHI2, DET, DF1, DF2
      REAL  L(3), LB(2,2), LBL(3), R(MXAUG), RX, TBAR(3)
      REAL  TIN(2,2,MXAUG), TZ(3)
      COMMON  /TINCOM/ TIN

      N = 1
      DF1 = REAL(NK)
      DF2 = REAL(NK)
      CALL RNCHI(MPT, DF1, CHIA)
      CALL RNCHI(MPT, DF2, CHIB)
      CALL RNNOR(MPT,R)

*     Begin DO WHILE loop
111   CONTINUE
      IF (N .GT. MPT) GOTO 999
*     Begin REPEAT UNTIL loop
222     continue
*         Calculate L, the lower Cholesky factor, note TIN = Tau_Star.
          L(1) = SQRT(TIN(1,1,N))
          L(2) = TIN(2,1,N) / L(1)
          L(3) = SQRT(TIN(2,2,N) - L(2) * L(2))
*         Set up B = AA', for A a Bartlett decomposition.
          B(1) = CHIA(N)
          B(2) = R(N) * SQRT(CHIA(N))
          B(3) = CHIB(N) + R(N) * R(N)
*         Calculate a new Tau = LBL / NK
          LB(1,1) = L(1) * B(1)
          LB(1,2) = L(1) * B(2)
          LB(2,1) = L(2) * B(1) + L(3) * B(2)
          LB(2,2) = L(2) * B(2) + L(3) * B(3)
          LBL(1) = LB(1,1) * L(1) /  10.0
          LBL(2) = (LB(1,1) * L(2) + LB(1,2) * L(3)) /  10.0
          LBL(3) = (LB(2,1) * L(2) + LB(2,2) * L(3)) /  10.0
*         Calculate TIN = Inverse(Tau)
          DET = LBL(1) * LBL(3) - LBL(2) * LBL(2)
          IF (DET .LT. 0.00001) THEN
            CALL RNCHI(1, DF1, CHI1)
            CALL RNCHI(1, DF2, CHI2)
            CALL RNNOR(1,RX)
            CHIA(N) = CHI1
            CHIB(N) = CHI2
            R(N) = RX
            TIN(1,1,N) = TZ(1)
            TIN(2,1,N) = TZ(2)
            TIN(2,2,N) = TZ(3)
            AGAIN = .TRUE.
          ELSE
            AGAIN = .FALSE.
          END IF
```

```
        IF (AGAIN) GOTO 222
*       End REPEAT UNTIL loop
        TIN(1,1,N) = LBL(3) / DET
        TIN(2,1,N) = -1.0 * LBL(2) / DET
        TIN(1,2,N) = TIN(2,1,N)
        TIN(2,2,N) = LBL(1) / DET
        N = N + 1
      GOTO 111
999   CONTINUE
*     End DO WHILE loop

*     Calculate point estimate of T and write it out.
      IF (IT .EQ. 25) THEN
          UN = 30
        DO 10 I = 1, 3
          TBAR(I) = 0.0
 10     CONTINUE
        DO 20 N = 1, MXAUG
          DET = TIN(1,1,N) * TIN(2,2,N) - TIN(1,2,N) * TIN(1,2,N)
          TBAR(1) = TBAR(1) + (TIN(2,2,N) / DET)
          TBAR(2) = TBAR(2) + (-1.0 * TIN(1,2,N) / DET)
          TBAR(3) = TBAR(3) + (TIN(1,1,N) / DET)
 20     CONTINUE
        DO 30 I = 1, 3
          TBAR(I) = TBAR(I) / REAL(MXAUG)
 30     CONTINUE
        WRITE(UN,301) (TBAR(J), J = 1, 3)
      END IF

301   FORMAT(3(F11.5))
      RETURN
      END
```

```
      SUBROUTINE BETA(BLS, ID2, MPT, NGEN, NK, WC, WE, WI, XX)
**********************************************************************
*     Calculate Beta_Star and then simulate the posterior distribution
*     of Beta.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  MPT, NGEN, NK, NR, NS, RPT
      REAL  B(2,MXK,MXAUG), BLS(2,MXK), BSTR(2), D(2,2), DET
      REAL  DIN(2,2), G(4,MXAUG), ID2(2,2), ILAM(2,2), ILAMW(2,4)
      REAL  ILAMWG(2), LAM(2,2), LF(3), R(20), SIG(MXAUG)
      REAL  TIN(2,2,MXAUG), WC(2,4), WE(2,4), WI(MXK), XX(2,2,MXK)
      COMMON  /BCOM/ B /GCOM/ G /SIGCOM/ SIG /TINCOM/ TIN

      NR = NK * 2
      NS = 1
      DO 10 N = 1, MPT
*       Get a set of random numbers.
        CALL RNNOR(NR, R)
        RPT = 1

        DO 20 K = 1, NK
*         Calculate DIN = Inv(Inv(V) + Inv(T)) for a group.
          DO 30 I = 1, 2
            DO 30 J = 1, 2
              D(I,J) = TIN(I,J,N) + XX(I,J,K) / SIG(NS)
 30       CONTINUE
          DET = D(1,1) * D(2,2) - D(1,2) * D(2,1)
          DIN(1,1) = D(2,2) / DET
          DIN(1,2) = -D(1,2) / DET
          DIN(2,1) = DIN(1,2)
          DIN(2,2) = D(1,1) / DET

*         Factor DIN = LF * LF' for generating a new Beta.
          LF(1) = SQRT(DIN(1,1))
          LF(2) = DIN(1,2) / LF(1)
          LF(3) = SQRT(DIN(2,2) - LF(2) * LF(2))

*         Calculate LAM = DIN * Inv(V)
          DO 40 J = 1, 2
            DO 40 I = 1, 2
              LAM(I,J) = 0.0
              DO 40 L = 1, 2
                LAM(I,J) = LAM(I,J) + DIN(I,L) * XX(L,J,K) / SIG(NS)
 40       CONTINUE

*         Calculate first part of Beta_Star = BSTR = LAM * BLS
          DO 50 I = 1, 2
            BSTR(I) = 0.0
            DO 50 J = 1, 2
              BSTR(I) = BSTR(I) + LAM(I,J) * BLS(J,K)
 50       CONTINUE

          DO 60 J = 1, 2
            DO 60 I = 1, 2
              ILAM(I,J)= ID2(I,J) - LAM(I,J)
 60       CONTINUE

*         Calculate the second part of Beta_Star = (I - LAM) * W * G.
```

```
          IF (WI(K) .GT. 0.0) THEN
            DO 70 J = 1, 4
              DO 70 I = 1, 2
                ILAMW(I,J) = 0.0
                DO 70 L = 1, 2
                  ILAMW(I,J) = ILAMW(I,J) + ILAM(I,L) * WE(L,J)
 70           CONTINUE
          ELSE
            DO 80 J = 1, 4
              DO 80 I = 1, 2
                ILAMW(I,J) = 0.0
                DO 80 L = 1, 2
                  ILAMW(I,J) = ILAMW(I,J) + ILAM(I,L) * WC(L,J)
 80           CONTINUE
          END IF

          DO 90 I = 1, 2
            ILAMWG(I) = 0.0
            DO 90 J = 1, 4
              ILAMWG(I) = ILAMWG(I) + ILAMW(I,J) * G(J,N)
 90       CONTINUE

*         Complete Beta_Star calculations.
          DO 100 I = 1, 2
            BSTR(I) = BSTR(I) + ILAMWG(I)
100       CONTINUE

*         Determine pointer value and generate new B's.
          B(1,K,N) = BSTR(1) + LF(1) * R(RPT)
          B(2,K,N) = BSTR(2) + LF(2) * R(RPT) + LF(3) * R(RPT+1)
          RPT = RPT + 2
 20     CONTINUE

*       Update pointer for Sigma_2.
        IF (NGEN .EQ. 2 .AND. MOD(N,2) .EQ. 0) THEN
          NS = NS + 1
        ELSE IF (NGEN .EQ. 1) THEN
          NS = N
        END IF
 10   CONTINUE

      RETURN
      END
```

```
      SUBROUTINE SSTAR(IT, MPT, NK, NT, SXY, SY2, XX)
*************************************************************************
*     Calculate the sum of squares for Sigma_2_Star for each data set.

      PARAMETER  (MXAUG=2048, MXK=10)
      INTEGER  IT, MPT, NK, NT, UN
      REAL  B(2,MXK,MXAUG), BXXB, SBAR, SCP, SIG(MXAUG), SS, SXY(2,MXK)
      REAL  SY2(MXK), XX(2,2,MXK), XXB(2)
      COMMON  /BCOM/ B /SIGCOM/ SIG

      DO 10 N = 1, MPT
        SS = 0.0
        DO 20 K = 1, NK
          DO 30 I = 1, 2
            XXB(I) = 0.0
            DO 40 J = 1, 2
              XXB(I) = XXB(I) + XX(I,J,K) * B(J,K,N)
 40         CONTINUE
 30       CONTINUE

          BXXB = 0.0
          SCP = 0.0
          DO 50 I = 1, 2
            BXXB = BXXB + B(I,K,N) * XXB(I)
            SCP = SCP + B(I,K,N) * SXY(I,K)
 50       CONTINUE
          SS = SS + SY2(K) - 2.0 * SCP + BXXB
 20     CONTINUE
*       SIG contains the sum of squares and has not been averaged.
        SIG(N) = SS
 10   CONTINUE

      IF (IT .EQ. 25) THEN
        UN = 30
        SBAR = 0.0
        DO 60 N = 1, MXAUG
          SBAR = SBAR + SIG(N) / REAL(NT)
 60     CONTINUE
        SBAR = SBAR / REAL(MXAUG)
        WRITE(UN,301) SBAR
      END IF
301   FORMAT(F11.5)

      RETURN
      END
```

```
      SUBROUTINE SIMS(IT, MPT, NT)
*************************************************************************
*     Simulate Sigma_2.  The vector SIG contains sum of squares for the
*     Sigma_2_Star's and these values are replaced with Sigma_2's.

      PARAMETER  (MXAUG=2048)
      INTEGER  IT, MPT, NT, UN
      REAL  CHI(MXAUG), DF, SBAR, SIG(MXAUG)
      COMMON  /SIGCOM/ SIG

      DF = REAL(NT)
      CALL RNCHI(MPT, DF, CHI)
      DO 10 N = 1, MPT
        SIG(N) = SIG(N) / CHI(N)
 10   CONTINUE

      IF (IT .EQ. 25) THEN
        UN = 30
        SBAR = 0.0
        DO 20 N = 1, MXAUG
          SBAR = SBAR + SIG(N)
 20     CONTINUE
        SBAR = SBAR / REAL(MXAUG)
        WRITE(UN,301) SBAR
      END IF

301   FORMAT(F11.5)

      RETURN
      END
```

```
      SUBROUTINE TAILS(IT, UN)
*************************************************************************
*     This subroutine calculates the median, upper and lower percentile
*     points of the posterior distribution of gamma for (alpha / 2) =
*     0.005, 0.025, 0.050.  TAILS is called by GSTAR.

      PARAMETER  (MXAUG=2048)
      INTEGER  IN, IT, MID, UN
      REAL  G(4,MXAUG), GVEC(MXAUG), P(7)
      COMMON  /GCOM/ G

*     Calculate the percentiles of the posterior distribution of Gamma.
      MID = MXAUG / 2
      DO 10 I = 1, 4
        DO 20 N = 1, MXAUG
          GVEC(N) = G(I,N)
 20     CONTINUE
        CALL SVRGN(MXAUG, GVEC, GVEC)
        P(1) = (GVEC(MID) + GVEC(MID+1)) / 2.0
        P(2) = GVEC(10) * 0.76 + GVEC(11) * 0.24
        P(3) = GVEC(2037) * 0.24 + GVEC(2038) * 0.76
        P(4) = GVEC(51) * 0.80 + GVEC(52) * 0.20
        P(5) = GVEC(1996) * 0.20 + GVEC(1997) * 0.80
        P(6) = GVEC(102) * 0.60 + GVEC(103) * 0.40
        P(7) = GVEC(1945) * 0.40 + GVEC(1946) * 0.60
        WRITE(UN,301) (P(J), J = 1, 7)
 10   CONTINUE

301   FORMAT(7(F11.5))

      RETURN
      END
```

```
      SUBROUTINE NTAVE(NEXP, NTOT)
************************************************************************
*     Calculate the average NT per experiment.

      INTEGER  NEXP, NTOT
      REAL  AVNT

      AVNT = REAL(NTOT) / REAL(NEXP)
      WRITE(40,401) AVNT

401   FORMAT(' Average N per experiment = ',F8.3)

      RETURN
      END
```

Hierarchical Data Generation FORTRAN Source Code Listing

```
      PROGRAM HLMDAT

************************************************************************
*     HLMDAT generates replications of an experiment.  The following
*     parameters must be set before each run.
*     NK - the number of classes
*     NEXP - the number of experiments
*     P0W - the correlation between B0j and Wj
*     P1W - the correlation between B1j and Wj
*     RNOPT(5) - initialization for IMSL number generator
*     RNSET(ISEED) - set the seed after the first run

      PARAMETER (MAX=22, NEXP=500)
      INTEGER  ISEED, NJ(MAX), NK
      DOUBLE PRECISION  G(4), P0W, P1W, SX(MAX), SXY(MAX), SX2(MAX)
      DOUBLE PRECISION  SY(MAX), SY2(MAX),T00, T00W, T11W, W(MAX)

      OPEN(10, FILE='SEED')
      OPEN(20, FILE='VARS')
      OPEN(30, FILE='DATA', FORM='UNFORMATTED')

*     Seed set for data set 3 with P0W = 0.0 and P1W = 0.6
      CALL RNOPT(5)
      ISEED = 1461009557
      CALL RNSET(ISEED)
      NK = 10

      P0W = 0.0
      P1W = 0.6
      CALL TGSET(G, P0W, P1W, T00, T00W, T11W)
      DO 10 K = 1, NEXP
        CALL CLSIZE (NJ, NK)
        CALL GEN(G, NJ, NK, SX, SXY, SX2, SY, SY2, T00W, T11W, W)
        CALL RITE(NEXP, NJ, NK, SX, SXY, SX2, SY, SY2, W)
 10   CONTINUE

      CALL RNGET(ISEED)
      WRITE(10,101) ISEED

101   FORMAT(I10)

      END
```

```
      SUBROUTINE TGSET(G, P0W, P1W, T00, T00W, T11W)
*************************************************************************
*     Set the between subject parameters Tau and Gamma.

      PARAMETER (C=0.10D+0, D=0.18D+0, PXY=0.60D+0)
      DOUBLE PRECISION  G(4), P0W, P1W, SIG2Y, S1
      DOUBLE PRECISION T00, T00W, T11, T11W

*     Calculate the unconditional and conditional variance terms
      T00 = D / ((1.0 - D) * (1.0 - PXY**2) - C * D)
      T11 = C * T00
      T00W = T00 * (1.0 - P0W**2)
      T11W = T11 * (1.0 - P1W**2)

*     The pooled unconditional variance of Y
*     S1 = C * T00 + 1.0
*     SIG2Y = S1 + (PXY**2 / (1.0 - PXY**2)) * (C * T00 + 1.0)

*     Calculate the Gamma vector G = {G00, G01, G10, G11}
      G(1) = 0.0
      G(2) = P0W * SQRT(T00)
      G(3) = SQRT((PXY**2 / (1 - PXY**2)) * (C * T00 + 1.0))
      G(4) = P1W * SQRT(T11)

      RETURN
      END
```

```
      SUBROUTINE CLSIZE (NJ, NK)
**********************************************************************
*     Randomly draw a vector of class sizes from a distribution.

      PARAMETER (MAX = 22)
      INTEGER  NJ(MAX), NK, CPT
      DOUBLE PRECISION  R(MAX)

      CALL DRNUN(NK, R)
      DO 10 I = 1, NK
        CPT = NINT(1000 * R(I))
        IF (CPT .GT. 995 .AND. CPT .LE. 1000) THEN
          NJ(I) = 30
        ELSE IF (CPT .GT. 965 .AND. CPT .LE. 995) THEN
          NJ(I) = 29
        ELSE IF (CPT .GT. 905 .AND. CPT .LE. 965) THEN
          NJ(I) = 28
        ELSE IF (CPT .GT. 795 .AND. CPT .LE. 905) THEN
          NJ(I) = 27
        ELSE IF (CPT .GT. 670 .AND. CPT .LE. 795) THEN
          NJ(I) = 26
        ELSE IF (CPT .GT. 535 .AND. CPT .LE. 670) THEN
          NJ(I) = 25
        ELSE IF (CPT .GT. 405 .AND. CPT .LE. 535) THEN
          NJ(I) = 24
        ELSE IF (CPT .GT. 280 .AND. CPT .LE. 405) THEN
          NJ(I) = 23
        ELSE IF (CPT .GT. 185 .AND. CPT .LE. 280) THEN
          NJ(I) = 22
        ELSE IF (CPT .GT. 115 .AND. CPT .LE. 185) THEN
          NJ(I) = 21
        ELSE IF (CPT .GT. 065 .AND. CPT .LE. 115) THEN
          NJ(I) = 20
        ELSE IF (CPT .GT. 035 .AND. CPT .LE. 065) THEN
          NJ(I) = 19
        ELSE IF (CPT .GT. 015 .AND. CPT .LE. 035) THEN
          NJ(I) = 18
        ELSE IF (CPT .GT. 005 .AND. CPT .LE. 015) THEN
          NJ(I) = 17
        ELSE
          NJ(I) = 16
        END IF
 10   CONTINUE

      RETURN
      END
```

```
      SUBROUTINE GEN(G, NJ, NK, SX, SXY, SX2, SY, SY2, T00W, T11W, W)
*****************************************************************
*     Generate data for one replication of an experiment.

      PARAMETER (MAX=22, MXR=61, MXU=44)
      INTEGER  NJ(MAX), NK, NR, NT
      DOUBLE PRECISION  B0(MAX), B1(MAX), CHI(2), DF1, DF2, G(4)
      DOUBLE PRECISION  R(MXR), SIG, SR(MAX), SR2(MAX), SX(MAX)
      DOUBLE PRECISION  SXR(MAX), SXSR(MAX), SXY(MAX), SX2(MAX), SY(MAX)
      DOUBLE PRECISION  SY2(MAX), TAU(3),TCHI, TMP1, TMP2, TMP3, T00W
      DOUBLE PRECISION  T11W, U(MXU), U0, U1, W(MAX)

      NT = 0
      SIG = 0.0
      DO 10 I = 1, 3
        TAU(I) = 0.0
10    CONTINUE

      DO 20 J = 1, NK
        NR = 2 * NJ(J)
        CALL DRNNOR(NR+1, R)
        NT = NT + NJ(J)
        SX(J) = 0.0
        SR(J) = 0.0
        SXR(J) = 0.0
        DO 30 I = 1, NR, 2
          SX(J) = SX(J) + R(I)
          SR(J) = SR(J) + R(I+1)
30      CONTINUE
        SXSR(J) = SX(J) * SR(J)
        DF1 = NJ(J) - 1
        CALL DRNCHI(2, DF1, CHI)
        SX2(J) = CHI(1) + SX(J)**2 / NJ(J)
        SR2(J) = CHI(2) + SR(J)**2 / NJ(J)
        SIG = SIG + SR2(J)
        DF2 = NJ(J) - 2
        CALL DRNCHI(1, DF2, TCHI)
        T = R(NR + 1) / SQRT(TCHI / DF2)
        P = T / SQRT(T**2 + DF2)
        SXR(J) =  P * SQRT(CHI(1)) * SQRT(CHI(2)) + SXSR(J) / NJ(J)
20    CONTINUE

      CALL DRNNOR(NK*2, U)
      DO 40 K = 1, NK
        IF (K .LE. NK/2) THEN
          W(K) = 1.0
        ELSE
          W(K) = -1.0
        END IF
        U0 = SQRT(T00W) * U(K)
        U1 = SQRT(T11W) * U(K+NK)
        TAU(1) = TAU(1) + U0 * U0
        TAU(2) = TAU(2) + U0 * U1
        TAU(3) = TAU(3) + U1 * U1
        B0(K) = G(1) + G(2) * W(K) + U0
        B1(K) = G(3) + G(4) * W(K) + U1
        SY(K) = NJ(K) * B0(K) + B1(K) * SX(K) + SR(K)
        TMP1 = NJ(K) * B0(K)**2 + B1(K)**2 * SX2(K) + SR2(K)
```

```
        TMP2 = 2.0 * B0(K) * B1(K) * SX(K) + 2.0 * B0(K) * SR(K)
        TMP3 = 2.0 * B1(K) * SXR(K)
        SY2(K) = TMP1 + TMP2 + TMP3
        SXY(K) = B0(K) * SX(K) + B1(K) * SX2(K) + SXR(K)
 40   CONTINUE

      SIG = SIG / NT
      DO 50 I = 1, 3
        TAU(I) = TAU(I) / NK
 50   CONTINUE
      WRITE(20,201) SIG
      WRITE(20,202) (TAU(I), I = 1, 3)

201   FORMAT(F11.5)
202   FORMAT(3F11.5)

      RETURN
      END
```

```
      SUBROUTINE RITE(NEXP, NJ, NK, SX, SXY, SX2, SY, SY2, W)
*************************************************************************
*     One replication of an experiment is written to a file.

      PARAMETER (MAX=22)
      CHARACTER *12  ID, LABEL1(0:2), LABEL2(0:1)
      CHARACTER *44  STRING
      INTEGER  M1, M2, NBR5, NEXP, NJ(MAX), NK, NVAR, QMAX, T, VNUM
      DOUBLE PRECISION  SX(MAX), SXY(MAX), SX2(MAX), SY(MAX)
      DOUBLE PRECISION  SY2(MAX), W(MAX), BAR(2)
      DATA STRING / '0102030405060708091011121314151617181920212 2'/

      NVAR = 2
      NBR5 = 0
      QMAX = 1
      VNUM = 0
      LABEL1(0) = '    BASE'
      LABEL1(1) = 'DEPY'
      LABEL1(2) = 'INDX'
      LABEL2(0) = '    BASE'
      LABEL2(1) = 'INDW'

      WRITE(30) NVAR, NBR5, QMAX, NK, VNUM
      WRITE(30) (LABEL1(I), I = 0, NVAR)
      WRITE(30) (LABEL2(I), I = 0, QMAX)
      M1 = 1
      M2 = 2
      DO 10 I=1, NK
        T = NJ(I)
        WRITE(30) T
        BAR(1) = SY(I) / T
        BAR(2) = SX(I) / T
        SY2(I) = SY2(I) - T * BAR(1) * BAR(1)
        SXY(I) = SXY(I) - T * BAR(1) * BAR(2)
        SX2(I) = SX2(I) - T * BAR(2) * BAR(2)
        WRITE(30) (BAR(K), K = 1, 2)
        WRITE(30) SY2(I), SXY(I), SX2(I)
        ID = STRING(M1:M2)
        WRITE(30) ID, W(I)
        M1 = M1 + 2
        M2 = M2 + 2
 10   CONTINUE

      RETURN
      END
```

REFERENCES

Anderson, T.W. (1984). *An introduction to multivariate statistical analysis.* New York: John Wiley and Sons.

Barcikowski, R.S. (1981). Statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 6*, 267-285.

Barr, R., and Dreeden, R. (1983). *How schools work.* Chicago: University of Chicago Press.

Bartlett, M.S. (1933). On the theory of statistical regression. *Proc. Roy. Soc. Edinburgh, 53*, 260-283.

Bassiri, D. (1988). *Large and small sample properties of maximum likelihood estimates for the hierarchical linear model.* Unpublished doctoral dissertation, Michigan State University.

Bates, D.M., and Watts, D.G. (1981). A relative offset orthoganality convergence criterion for nonlinear least squares. *Technometrics, 23, 2*, 179-184.

Bayes, T.R. (1763). An essay towards solving a problem in the doctrine of chances. *The Philosophical Transactions of the Royal Society, 53*, 370-418 (reprinted in *Biometrika* (1958), 45, 293-315).

Bidwell, C. and Kasarda, J. (1980). Conceptualizing and measuring the effects of school and schooling. *American Journal of Education, 88*, 401-430.

Blair, R.C., and Higgins, J.J. (1986). Comment on statistical power with group mean as the unit of analysis. *Journal of Educational Statistics, 11*, 161-169.

Blair, R.C., Higgins, J.J., Topping, M.E.H., and Mortimer, A.L. (1983). An investigation of the robustness of the t test to unit of analysis violations. *Educational and Psychological Measurement, 43*, 69-80.

Box, G.E.P., and Tiao, G.C. (1973). *Bayesian inference in statistical analysis.* Reading, Massachusetts: Addison-Wesley Publishing Company.

Braun, H.I., Jones, D.H., Robin, D.B., and Thayer, D.T. (1983). Empirical Bayes estimation in the general linear model with data of deficient rank. *Psychometrika, 48, 2,* 171-181.

Brophy, J.E., and Good, T.L. (1986). Teacher behavior and student achievement. In M.C. Wittrock (Ed.), *Handbook of research on teaching.* New York: Macmillan.

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, R.T. (1986). *An introduction to HLM: Computer program and users' guide.* University of Chicago, Department of Education.

Burden, R.L., Faires, J.D., and Reynolds, A.C. (1981). *Numerical analysis.* Boston: Prindle, Weber and Schmidt.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. In D.C. Berliner (Ed.), *Review of research in education, vol. 8.* Washington, DC: American Educational Research Association.

Campbell, D.T., and Stanley, J.C. (1963). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally College Publishing Company.

Chambers, J.M. (1977). *Computational methods for data analysis.* New York: John Wiley & Sons.

Cooley, W.W., Bond, L., and Mao, B. (1981). Analyzing multilevel data. In Berk, R.A. (Ed.), *Educational evaluation methodology.* Baltimore: Johns Hopkins University Press.

Cowles, M. and Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist, 37,* 553-558.

Cronbach, L.J. (1957). The two disciplines of scientific psychology. *The American Psychologist, 12,* 671-684.

Cronbach, L.J., and Snow, R.E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Cronbach, L.J., and Webb, N. (1975). Between and within-class effects in a reported aptitude-by-treatment interaction: Reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology, 6,* 717-724.

DeLeeuw, J., and Kreft, I. (1986). Random coefficient models for multilevel analysis. *Journal of Educational Statistics, 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*, 1-38.

Dempster, A.P., Rubin, D.B., and Tsutakawa, R.K. (1981). Estimation in covariance components models. *Journal of the American Statistical Association, 76*, 341-353.

Edwards, A.L. (1948). Note on the "correction for continuity" in testing the significance of the difference between correlated proportions. *Psychometrika, 13*, 185-187.

Fisher, R.A. (1936). *Statistical methods for research workers*. Edinburgh: Oliver and Boyd.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley & Sons.

Glass, G.V., and Stanley, J.C. (1970). *Statistical methods in education and psychology*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Goldstein, H.I. (1987). *Multilevel models in education and social research*. London: Oxford University Press.

Good, T.L., and Brophy, J.E. (1986). School effects. In M.C. Wittrock (Ed.), *Handbook of research on teaching*. New York: Micmillan.

Harville, D. A. (1976). Extension of the Gauss-Markov Theorem to include the estimation of random effects. *Annals of Statistics, 4*, 384-95.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association, 72*, 320-338.

Henderson, C.R. (1953). Estimation of variance and covariance components. *Biometrics, 9*, 226-252.

Henderson, C.R., Jr., and Henderson, C.R. (1979). Analysis of covariance in mixed models with unequal subclass numbers. *Communications in Statistics-Theory and Methods*, 751-787.

Heyns, B. (1986). Educational effects: Issues in conceptualization and measurement. In J.G. Richardson (Ed.), *Handbook of theory and research for the sociology of education*. Westport, CT: Greenwood Press.

Hopkins, K.D. (1982). The unit of analysis: Group means versus individual observations. *American Educational Research Journal, 19*, 5-18.

International Mathematical and Statistics Libraries (1987). *MATH/LIBRARY and STAT/LIBRARY, version 1.0.* Houston: IMSL Problem-Solving Software Systems.

Jones, M.C. (1985). Generating inverse Wishart matrices. *Communications in Statistics-Simulation and Computation*, 511-514.

Johnson, R.A., and Bhattacharyya, G.K. (1977). *Statistical concepts and methods.* New York: John Wiley and Sons.

Kackar, R.N., and Harville, D.A. (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association, 79*, 853-862.

Kirk, R.E. (1968). *Experimental design: Procedures for the behavioral sciences.* Belmont, California: Brooks/Cole Publishing Company.

Kish, L. (1965). *Survey sampling.* New York: Wiley and Sons.

Knuth, D.E. (1981). *The art of computer programming, volume 2, seminumerical algorithms.* Reading, Massachusetts: Addison-Wesley Publishing Company.

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

Lindley, D.V. (1965a). *Introduction to probability and statistics from a Bayesian viewpoint, part 1, probability.* Cambridge: Cambridge University Press.

Lindley, D.V. (1965b). *Introduction to probability and statistics from a Bayesian viewpoint, part 2, inference.* Cambridge: Cambridge University Press.

Lindley, D.V., and Smith, A.F.M. (1972). Bayes estimates for the linear model (with discussion). *Journal of the Royal Statistical Society, Series B, 34*, 1-41.

Lindquist, E.F. (1953). *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin.

Lumsdaine, A.A. (1963). Instruments and media of instruction. In N.L. Gage (Ed.), *Handbook of research on teaching.* Chicago: Rand-McNally.

Longford, N.T., (1987). A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested effects. *Biometrika, 74*, 817-827.

Mason, W.M., Wong, G.Y., and Entwistle, B. (1984). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72-103). San Francisco: Jossey-Bass.

Miettinen, O.S. (1968). The matched pairs design in the case of all-or-none responses. *Biometrics, 24*, 339-352.

Morris, C.N. (1987). Comment on Tanner and Wong (1987). *Journal of the American Statistical Association, 82*, 542-543.

Morrison, D.F. (1976). *Multivariate statistical methods*. New York: McGraw-Hill.

Odel, P.L., and Feiveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *Journal of the American Statistical Association, 61*, 199-203.

Peckman, P.D., Glass, G.V., and Hopkins, K.D. (1969). The experimental unit in statistical analysis. *Journal of Special Education, 3*, 337-349.

Pedhazur, E.J. (1982). *Multiple regression in behavioral research: Explanation and prediction.* New York: Holt, Rinehart and Winston.

Raudenbush, S.W. (1984). *Applications of a hierarchical linear model in educational research.* Unpublished doctoral dissertation, Harvard University.

Raudenbush, S.W., and Bryk, A.S. (1985). 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*, 47-65.

Raudenbush, S.W., and Bryk, A.S. (1988). Methodological advances in analyzing the effects of schools and classrooms on student learning. In E.Z. Rothkopf (Ed.), *Review of research in education* (pp 423-475). Washington, DC: American Educational Research Association.

Searle, S.R. (1970). Large sample variances of maximum likelihood estimators of variance components using unbalanced data. *Biometrics, 26*, 505-524.

Searle, S.R. (1971). Topics in variance component estimation. *Biometrics, 27*, 1-76.

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

Smith, W.B., and Hocking, R.R. (1972). Algorithm AS53: Wishart variate generator. *Applied Statistics, 21*, 341-345.

Strenio, J.F., Weisberg, H.I., and Bryk, A.S. (1983). Empirical Bayes estimation of individual growth-curve parameters and their relationship to covariates. *Biometrics, 39*, 71-86

Tankard, J.W., Jr. (1984). *The statistical pioneers*. Cambridge, Massachusetts: Schenkmann Publishing Company, Inc.

Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *Journal of the American Statistical Association, 82*, 528-550.

Thisted, R.A. (1988). *Elements of statistical computing*. New York: Chapman and Hall.

Winer, B.J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill Book Company.

Wu, C.F.J (1983). On the convergence properties of the EM algorithm. *Annals of Statistics, 11*, 95-103.