THESIS

MICHIGAN STATE UNIVERSITY LIBRARIES

3 1293 01020 4976

This is to certify that the

dissertation entitled

A MULTIVARIATE MIXED LINEAR MODEL
FOR META ANALYSIS

presented by

HRIPSIME A. KALAIAN

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Education

Major professor

Date August 4, 1994

MSU is an Affirmative Action/Equal Opportunity Institution

O-12771

LIBRARY
Michigan State
University

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.

DATE DUE	DATE DUE	DATE DUE
FEB 04 1998	JUN 06 2005	
MAGIC 2 DEC 28 1998		
NOV 16 2002		

MSU Is An Affirmative Action/Equal Opportunity Institution
c:\circ\datedue.pm3-p.1

A MULTIVARIATE MIXED LINEAR MODEL FOR META-ANALYSIS

By

Hripsime A. Kalaian

A DISSERTATION

Submitted to
Michigan State University
College of Education
In partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational Psychology,
and
Special Education

1994

ABSTRACT

A MULTIVARIATE MIXED LINEAR MODEL FOR META-ANALYSIS

By

Hripsime A. Kalaian

Meta-analysts often encounter data sets with multiple effect sizes from each primary study in the review either because of multiple measures or multiple treatments. Having these correlated multiple effect sizes requires the use of multivariate analytical techniques which take into account the intercorrelations among these multiple effect sizes.

In the present study, the multivariate mixed-effects model for meta-analysis is developed and presented. This multivariate model takes into account three important characteristics which often arise in meta-analysis. The first is having multiple correlated effect sizes. The second is that different studies can have different subsets of effect sizes depending on the design of the primary study. The third is that these multiple effect sizes may be random realizations from a population of possible effect sizes. Using the

proposed model enables meta-analysts to obtain multivariate empirical Bayes estimates of the parameters in the model without excluding studies when some of the effect sizes are missing.

The application of the multivariate mixed-effects model is illustrated using multivariate artificial effect sizes (generated from the multivariate normal distribution) and a real data set. The real data set involves Scholastic Aptitude Test (SAT) coaching studies evaluating the effects of coaching on the two SAT subtests (SAT-Verbal and SAT-Math). Also, the fixed-effects model parameter estimates obtained from analyzing the transformed GLS model are compared to the mixed-effects model parameter estimates obtained from the HLM program.

In conclusion, the multivariate mixed-effects model using the HLM program can be applied to multivariate meta-analysis studies with missing effect sizes to obtain empirical Bayes estimates. Also, the proposed model can be used to perform multivaraite fixed-effects analysis. Finally, the findings of the present study can be generalized to studies with more than two outcomes (effect sizes) and at the same time within-study characteristics can be incorporated in these applications.

ACKNOWLEDGEMENTS

This study would not have been possible without the help of many individuals. My appreciation is offered to them for their encouragements and support.

In one way or another, each of the members of my doctoral committee has shared the decade with me. First, I would like to thank Dr. Steve Raudenbush, my advisor and dissertation chair, for his constant support and valuable advice. I thank him and his family for helping us (me and my family) through our life crisis by providing a loving friendship. Second, I would like to thank Dr. Betsy Becker for her belief in my work. I thank her for valuing me by listening and for responding seriously and quickly to my educational and personal problems. Third, I would like to thank Dr. Richard Houang for his constant help, support, and advice through my study. Finally, I would like to thank Dr. Dennis Gilliland for being the best teacher. He taught me not only statistical subjects, but how to deal with students and treat them with respect.

I would also like to express my deepest gratitude to my

husband, Rafa, for his support and help in any way he can to make this goal attainable. Our Wonderful three children, Nader, Neda, and Nabeel deserve special thanks for their sacrifices and patience so I can finish my studies.

I also want to thank my colleagues in the Office of Medical Education who have upheld me and this study in both professional and practical ways. Dr. Bob Bridgham, Dr. Patricia Mullan, Dr. Andrew Hogan, Dr. Rebecca Henry, and Mrs. Karen Boatman all in one way or another supported and helped me to achieve my goals.

Finally, I am deeply indebted to both of my parents and my sisters and brother, each of whom has prayed for me, supported me in every aspect of my life, taught me to value education and hard work, and loved me through this experience.

TABLE OF CONTENTS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

1. META-ANALYSIS IN EDUCATIONAL AND SOCIAL SCIENCES

In the last two decades there has been a surge of interest among educational and social researchers in applying quantitative methods for synthesizing and aggregating the results of primary related studies. The goals of research synthesis are accumulating and combining research evidence from many studies testing the same research hypothesis and also generating new evidence which helps to formulate new research hypotheses and plan future research studies. In other words, meta-analysis is, potentially, a powerful tool for synthesizing existing knowledge, criticizing the design of existing research, and stimulating more meaningful interdisciplinary research.

Various quantitative methods for research synthesis have been developed and applied within the last twenty years. One way of synthesizing and summarizing the research findings from

previous investigations is by aggregating effect magnitudes using meta-analysis statistical techniques. The term "meta-analysis" was first introduced and popularized to the social science literature by Glass (1976), and has also been developed by others, such as Rosenthal (1978) and Rosenthal and Rubin (1979). Pillemer and Light (1980) and Cooper (1982) provided a conceptual framework for research synthesis. Cooper (1982, 1984) developed a systematic approach (five-stage model) to carry on a research synthesis and an integrative research review. Hedges (1981, 1982, 1983), and Hedges and Olkin (1985) introduced the technical statistical methods for meta-analysis. Rosenthal (1978) presented a collection of statistical procedure for combining significance levels from primary research.

Meta-analysis can be defined as the statistical analysis of a large collection of primary research studies which focus on the same research question for the purpose of accumulating previous findings and consequently generating new research evidence. The most popular meta-analysis technique is first calculating an effect size for each primary study in the sample of collected studies in the review and then finding an overall effect-size estimate (here we assume that the effect sizes from the primary studies share a common population effect size). Thus, for treatment-control studies, effect size can be defined as the standardized mean difference between the experimental and control groups from each study in

the integrative review.

2. META-ANALYSIS IN MEDICAL SCIENCES

Since the mid-1980s the application of meta-analysis techniques for research review purposes spread from social and behavioral sciences through many other disciplines, especially medical sciences and health care disciplines. Meta-analyses of clinical trials (e.g., Yusuf et. al., 1987; Havens et. al, 1988) and epidemiologic studies (e.g. Longnecker et. al, 1988; Shinton and Beevers, 1989; Berlin and Colditz, 1990; Greenland, 1993) have been used frequently as an attempt to improve on traditional methods of narrative review. As in educational and behavioral sciences, the aim of the meta-analysis in health-care disciplines is systematically aggregating and summarizing data from the primary clinical trial studies to obtain a quantitative estimate of the overall effect of a particular treatment or clinical procedure on a defined outcome. Many meta-analysts have reviewed and examined the methodology of meta-analysis as applied to clinical problems especially to randomized controlled trials (Ottenbacher and Petersen, 1983; DerSimonian and Laird, 1986; L'Abbee', Detsky, and O'Rourke, 1987; Sacks et. al, 1987; Jenicek, 1989; Thacker, 1988; Greenland, 1987). Gerberg and

Horwitz (1988) presented guidelines for conducting meta-analysis for clinical studies. Huque (1988) defines meta-analysis as a statistical analysis which combines or integrates the results of several independent clinical trials considered by the meta analyst to be integrable.

3. MULTIPLE DEPENDENT EFFECT SIZES

Educational and social researchers often try to examine and explain a behavioral phenomenon by collecting multiple measurements from each individual in the study. As a result of having multiple measurements, primary research studies are not always so simple to integrate and summarize. Thus, meta-analysts usually calculate multiple measures for the effect of the experimental treatment depending on the number of the outcome variables in each study in the review.

Some of these studies compare different treatment groups to a single control group and are called <u>multiple treatments studies</u>. Other studies compare a single treatment group to a single control group, but instead of obtaining a single outcome measure, multiple outcome measures are obtained where there are several subscales in the outcome measure or test. These will be referred to as <u>multiple measures studies</u>.

Moreover, another set of studies, which can be characterized as pretest-posttest study designs, compare a single treatment group to a single control group and multiple pretest and posttest outcome measures are obtained from each study. These type of studies are referred to as <u>pre-post multiple measures studies</u>.

4. MULTIVARIATE STATISTICS

Having these correlated multiple effect magnitudes from each primary study in the review requires multivariate procedures of analysis (Hedges & Olkin, 1985; Raudenbush, Becker & Kalaian, 1988). Multivariate analysis refers to a collection of descriptive and inferential methods that have been developed for situations where we have more than one outcome variable and these outcome variables are correlated. Using multivariate procedures for analyzing meta-analysis data sets with multivariate characterization has various advantages. For example, (a) it provides us with better parameter estimates because it handles the multiple effect sizes simultaneously, taking into account the interdependence among the outcome variables, (b) it controls Type I error rates, (c) it also facilitate statistical comparisons among outcomes.

5. PURPOSE OF THE PRESENT STUDY

This thesis will present a multivariate mixed-effects model (multivariate hierarchical linear model) for meta-analysis that considers the multiple effect sizes from multiple-outcome studies or multiple-treatment studies from each study as random, and then models these effect sizes or the correlation coefficients as a function of study characteristics plus random error. Thus, this multivariate model takes into account three important characteristics of this type of data which often arise in meta-analysis. The first is having multiple effect sizes based on multiple dependent variables from each study. The second important characteristic is that different studies can have different subsets of dependent variables and consequently different numbers of effect sizes and correlations for each study. The third characteristic is that the effect sizes and the product-moment correlation coefficients from several studies are often viewed as random realizations from a population of possible effect sizes and correlation coefficients.

The application of the proposed multivariate mixed-effects model will be evaluated and examined empirically using artificial and real data sets. The artificial multiple effect sizes will be generated from the multivariate normal

distribution with specified mean vector and variance-covariance matrix. These effect sizes will be analyzed and compared by using the Hierarchical Linear Model (HLM) program (designed for analyzing multi-level data) and the V-Known routine (designed for meta-analysis purposes when the within-study variance-covariance matrices are known).

The real data set represents the Scholastic Aptitude Test (SAT) coaching studies. These multiple effect sizes represent the effects of coaching on SAT-Verbal and SAT-Math scores. These effect sizes will be evaluated by using the HLM program.

6. ADVANTAGES OF USING MULTIVARIATE MIXED MODEL

The estimates and hypothesis-testing procedures generated by using the multivariate mixed-linear model are fully multivariate techniques since they take into account the correlations among the multiple effect sizes from each study and meanwhile have several important properties. They allow one:

1. To distinguish between variation in the true multiple effect size parameters for each study, and the sampling covariation which results because effect sizes are

estimated with error. That is

$$\text{Total Effect Size Covariation} = \text{Parameter Covariation} + \text{Error Covariation}$$

2. To examine the differential effects of the treatment on the multiple outcome measures;

3. To test hypotheses about the effects of study characteristics and features on multiple study outcomes;

4. To estimate the variance-covariance matrix of the multiple random effects and test the hypothesis of no variation-covariation among the multiple effect size parameters;

5. To find improved empirical Bayes estimates of multiple effect sizes and multiple product-moment correlation coefficients in each study;

6. To include in the analysis different numbers of outcomes from each study as well as different predictors for the different outcome measures;

7. To provide more precise and stable parameter estimates.

7. ORGANIZATION OF THE PRESENT STUDY

This study contains eight chapters dealing with the theory and the application of the multivariate mixed-effects model for meta-analysis and research integration. Chapter two will review the existing literature on the statistical approaches and methods of meta-analysis.

Chapter 3 will present a description of the notation and the statistical terms used for the multivariate hierarchical linear model. Also, the theoretical background and notation for meta-analysis will be reviewed in this chapter.

The multivariate mixed-effects model for meta-analysis will be introduced and developed in Chapter 4. First, the unconditional model (with no predictor in the model) will be illustrated. Second, the conditional model (where the variations among the multiple effect sizes are explained by some study predictors) will be explained.

Chapter 5 will deal with the estimation of the multivariate mixed-effects model that proposed in this study. Also, the maximum likelihood method of estimation and the EM algorithm will be presented in order to obtain empirical Bayes estimates of the parameters in the model.

In Chapter 6, an artificial multivariate effect-size data set will be generated using FORTRAN and IMSL subroutines. The

results of applying the proposed model to these generated data using the HLM program for analyzing multi-level data and the V-Known routine for analyzing effect-size data will be compared. The findings of this chapter will help us to pursue the use of the HLM program for meta-analysis purposes, especially when there are missing effect sizes in the data set.

Chapter 7 will present empirical results of applying the proposed multivariate mixed-effects model to Scholastic Aptitude Test (SAT) coaching data. The results and the conclusions based on fitting unconditional and conditional hierarchical linear models will be documented. Also, in this chapter, the applicability of the proposed multivariate mixed-effects model to obtain multivariate fixed-effects parameter estimates of the effects of the SAT coaching will be illustrated and these parameter estimates will be compared to those estimates from the multivariate mixed-effects model.

Finally, in Chapter 8, a concluding statement on the results of applying the proposed model to the artificial generated data and the SAT coaching studies will be presented. Also, the implications of the findings for further research related to multivariate effect-size meta-analysis will be discussed.

CHAPTER II

REVIEW OF THE LITERATURE

There has been much research and development progress in meta-analysis techniques in the last two decades. The developments have included tests of homogeneity of the effect sizes, modeling heterogeneity using fixed-effects and random-effects models for univariate effect sizes and correlation coefficients, and modeling multivariate effect sizes for fixed-effects cases. In this chapter the statistical techniques used previously to analyze data from studies that have multiple outcome measures are reviewed.

1. UNIVARIATE APPROACHES

Despite the multivariate characterization of the situations of multiple outcome variables from each study, the most frequently used procedure is to treat the multiple effect

sizes separately, with one meta-analysis for each outcome measure (e. g., Giaconia & Hedges, 1982; Kulik & Kulik, 1984; Rosenthal & Rubin, 1978; White, 1976). This practice of dealing with multiple outcome effect sizes and correlation coefficients individually inflates Type I error rates for quantitative review results, which in turn decreases the future replicability of the research findings. Moreover, conducting a separate meta-analysis for each outcome measure limits the kinds of research questions that the meta analyst can address. For example, the research questions 'Does a specific treatment have differential effects on the multiple outcomes?' or 'Does a specific study characteristic have differential effects on the multiple product-moment correlation coefficients?' cannot be answered precisely and accurately using univariate meta-analysis procedures.

Another common method of meta-analysis is to combine the estimates of the multiple effect sizes such as by averaging or summing the effect sizes for the multiple outcomes or the multiple correlation coefficients (e. g., Iaffaldano & Muchinsky, 1985). Employing this pooling procedure may result in losing important information about variation between the multiple effect sizes because a single treatment may have different effects on different outcome measures. This procedure is more appropriate when the outcomes represent or measure the same construct. Hedges and Olkin (1985) proposed a test for homogeneity of multiple effect sizes within each

study and a pooling procedure under the assumption that the multiple outcomes are measures of a single construct. Univariate statistical theories for synthesizing research studies are described below.

1.1 Univariate Fixed-Effects

This approach stresses the estimation of a fixed and common population effect of the treatment across a series of studies which test the same research hypothesis (Glass, 1976; Hedges, 1981). The method involves the calculation of an estimate of effect size from each single study. The average of effect-size estimates across studies for each outcome measure is used as an index of the overall effect size for each of the multiple outcome measures. Hedges (1982a) developed a test of homogeneity of effect-size estimates. This test examines whether the observed effect-size estimates vary by more than would be expected if all studies shared a common underlying population effect sizes.

Further, if the test of homogeneity fails, the meta-analyst tries to construct a weighted least squares regression model or a categorical model by regressing effect size estimates on various known study features (Hedges, 1982b). The main reason to use a regression model is to explain the

variability among the effect-size estimates from different studies by using known study characteristics as predictors.

1.2 Univariate Random-Effects

Contrary to the fixed-effects model, which assumes that there is a single underlying population effect of the treatment across all studies or that all the variation between studies can be explained by known study characteristics, the random-effects model assumes that the values of the effect sizes are sampled from a distribution of effect-size parameters. In other words, in the random-effects model there is no single true population effect. The true effects are from a distribution of effects.

Thus, by using the random-effects models, we can estimate the variance components of the distribution of the population of effect-size parameters as well as the variance components of the sampling distribution of the effect sizes. In other words, there are two sources of variation in the observed effect sizes (variability in the population effect-size parameter distribution and the variability in the effect-size estimates about the true parameter values.

Rubin (1981) suggested a random-effects model to summarize the results from parallel randomized experiments.

He used Bayesian and empirical Bayesian techniques to obtain improved estimates of the treatment effects in each experiment. Thus, his model views study effects as being random realizations of a population of treatment effects. Moreover, this model enables the researcher to estimate the variance of the treatment effect parameters. However, since the parallel randomized experiments have the same outcome measure, he did not incorporate the standardized effect-size estimates in his model. Also, he did not model the variation among the parallel experiments as a function of experiment characteristics.

DerSimonian and Laird (1983) used the univariate random effects model in their meta-analysis to estimate an overall average effect of SAT coaching. Also, they obtained empirical Bayes estimates of the individual study and program effects as well as their estimated variances via the EM algorithm using the maximum likelihood estimation procedure. Their outcome was not the effect size, $\underline{d}$, rather they looked at raw mean differences.

Hedges (1983) developed the statistical theory for the random-effects model for effect sizes. In this model the effect sizes are not assumed fixed but instead are viewed as sample realizations from a distribution of possible population effect size parameters with a mean and variance to be estimated via methods of moments. Thus, by using this model, the observed variance among treatment effects can be

decomposed into two components (a) sampling error or conditional variability of the estimated effect sizes around its population effect sizes and (b) random variation of the individual study effect sizes around the mean population effect size.

1.3 Univariate Mixed-Effects

The mixed-effects model corresponds to a setup with both fixed and random treatment effects. The random effects are the residuals (effect parameters minus predicted values) and the fixed effects are the effects of between study predictors.

Raudenbush and Bryk (1985), building on the work of Rubin, provided a statistical theory for a univariate hierarchical linear model (mixed-effects model) for meta-analysis. Their model views the effect sizes are random and models the variation among the effect sizes as a function of study characteristics plus error. Also, their model enables the meta-analyst to find improved empirical Bayes estimates of individual effect sizes.

Raudenbush (1988) reformulated the hierarchical linear model as the general mixed-model. This model allows estimation of the random and fixed effects when the within-group predictor matrices are less than full rank.

2. MULTIVARIATE APPROACHES

We characterize a procedure as being "multivariate" when we have multiple effect sizes on the basis of having multiple dependent measures or multiple treatment groups compared to a common control group for each study. Consequently, we analyze this kind of data simultaneously by taking into account the intercorrelations among the multiple outcomes or the multiple treatments. That is, we consider a procedure as being multivariate where several measurements or treatments are modeled jointly.

2.1 Multivariate Fixed-Effects

Hedges and Olkin (1985) proposed a multivariate statistical theory for summarizing the results from different studies with multiple outcome measures. Their approach requires that all studies use the same number of outcome measures. However, they didn't provide a statistical model to explain the variability in multiple effect sizes as a function of study features and experimental conditions.

Rosenthal and Rubin (1986) presented another method for

combining and comparing research results from studies having multiple effect sizes based on multiple dependent variables. They provided a method for obtaining a single summary effect size estimate from multiple effect sizes and a technique for testing this composite effect size. Also, they described a procedure for estimating the magnitude of the effect for a contrast among the multiple effect sizes of an individual study and for testing the significance of this contrast effect size. Their proposed meta-analytic procedures do not allow different predictors for the various dependent variables. They also did not provide a model to explain the variability in multiple effect sizes as a function of study characteristics.

Raudenbush, Becker, and Kalaian (1988) proposed generalized least squares (GLS) regression to model the variation between studies and to account for the interdependence among multiple outcomes within studies. Their approach allows the meta-analyst to include in the analysis different numbers of outcome measures from each study and different sets of predictors for each outcome measure. They view study effects as fixed, which means that all the variation among the multiple study effects other than sampling variance and covariance can be explained as a function of study characteristics.

3. SUMMARY OF PREVIOUS META-ANALYSIS TECHNIQUES

Four main techniques have been used previously to deal with studies that have multiple outcomes and consequently multiple effect sizes. The first and the most commonly used approach is the univariate fixed-effects model where the meta-analyst conducts a separate meta-analysis for each outcome measure. The basic assumption of this model is that the treatment and control populations share a common effect size, and the existing differences among these effect sizes can be determined through the knowledge of some study characteristics (Glass, 1976; Hedges, 1981). The univariate random-effects model is the second approach where the investigator also deals with the multiple outcomes separately. By using this approach the researcher assumes that there is a distribution of true effects for the experimental and control populations (Rubin, 1981; Hedges, 1983).

The third approach is the univariate mixed-effects approach (Raudenbush & Bryk, 1985; Raudenbush, 1988) where the estimated effect sizes can be modeled as a function of study characteristics plus random error. These univariate approaches all assume that multiple outcomes from each study are independent.

The fourth approach is the multivariate fixed-effects model (Raudenbush, Becker & Kalaian, 1988; Gleser & Olkin,

1993) which assumes that the study effects are fixed and considers all the variation-covariation among the standardized multiple study effects other than sampling variances and covariances to be explainable as a function of study characteristics (study design, treatment conditions, contexts, etc.).

In summary, these previous meta-analysis techniques either didn't account for the intercorrelations between the multiple outcome measures (univariate procedures) or assumed that the size of the multiple effects reported in each study depend strictly on known study characteristics and all of the variation between these studies can be explained by these known predictors.

Table 1

Previous Meta-analysis approaches for effect size data

	Fixed-Effects	Random-and-Mixed-Effects
Univariate approaches	Glass (1976) Hedges (1981)	Rubin (1981) Hedges (1983) Dersimonian & Laird (1983) Raudenbush & Bryk (1985)
Multivariate approaches	Hedges & Olkin (1985) Rosenthal & Rubin (1986) Raudenbush, Becker, & Kalaian (1988) Gleser & Olkin (1994)	

CHAPTER III

NOTATION FOR MULTIVARIATE MIXED LINEAR MODEL

Here we should distinguish between three kinds of studies, multiple measures, multiple treatments, and pre-post multiple measures studies. In multiple measures studies a single treatment group is compared to a single control group in each study and multiple outcome measures are obtained from each study. On the other hand, in multiple-treatments studies, multiple treatment groups are compared to a common control group in each study on a single outcome variable or multiple treatment group means are contrasted in each study.

As in multiple measures studies, in the third kind of study, a single treatment group is compared to a single control group in each pretest-posttest study and multiple pretest and posttest outcome measures are obtained from each study. This differentiation is made because (a) the estimated effect sizes and their variances for pre-post study designs are different from the other two kinds of studies, and (b) the formulas for estimating the covariances between the estimated effect sizes are different for the three kinds of studies. Thus, each type of study must be separately considered.

1. MULTIPLE MEASURES FOR EACH STUDY

The model for multivariate mixed meta-analysis for multiple measure studies assumes that we have K studies each comparing an experimental treatment (E) to a control condition (C) on one or more of p_i outcome measures (in study i). Where $i = 1,2,\ldots,K$ studies.

Let the outcome measures Y_{ijp}^{E} and Y_{ijp}^{C} for person j on outcome p in study i be normally distributed with means μ_{ip}^{E} and μ_{ip}^{C}, respectively and with common variance σ_{ip}^{2}. Thus, we assume that

$$Y_{ijp}^{E} \sim N(\mu_{ip}^{E}, \sigma_{ip}^{2}),$$
$$Y_{ijp}^{C} \sim N(\mu_{ip}^{C}, \sigma_{ip}^{2}).$$

where,

$j = 1,2,\ldots,n_i^{E}$ subjects, or $j = 1,2,\ldots,n_i^{C}$ subjects,

$i = 1,2,\ldots,K$ studies, and

$p = 1,2,....,P_i$ outcome measures.

1.1 Glass's Estimate of Effect Size

Glass (1976) proposed that the standardized mean difference between the experimental and control groups for the pth outcome measure, Y_{ip}, in the ith study is

$$g_{ip} = \frac{\overline{Y}_{ip}^{E} - \overline{Y}_{ip}^{C}}{S_{ip}} ,$$

where $\overline{Y}_{ip}^{E}$ and $\overline{Y}_{ip}^{C}$ are the ith experimental and control group means respectively for the pth outcome measure, Y_{ip}. Also S_{ip}^2 is the pooled within-groups estimate of the sample variance which can be calculated as

$$S_{ip}^2 = \frac{(n_i^E - 1)(S_{ip}^E)^2 + (n_i^C - 1)(S_{ip}^C)^2}{n_i^E + n_i^C - 2} ,$$

where S_{ip}^{E} and S_{ip}^{C} are the experimental and control group standard deviations, respectively.

1.2 Population Effect Size

Hedges (1981) developed the distribution theory for the effect size. He indicated that g_{ip} estimates a population effect size for the pth outcome measure for the ith study. The parameter δ_{ip} can be represented as

$$\delta_{ip} = \frac{\mu_{ip}^{E} - \mu_{ip}^{C}}{\sigma_{ip}} ,$$

where σ_{ip} is the pooled within-groups population standard deviation and μ_{ip}^{E} and μ_{ip}^{C} are the ith experimental and control population means for the pth outcome measure, respectively.

1.3 Unbiased Estimate of Effect Size

Hedges (1981) also indicated that Glass's estimator g_{ip} is a biased estimator of the population effect-size δ_{ip} and he derived the minimum variance unbiased estimator, d_{ip}, which is approximately

$$d_{ip} = c(m_i)\, g_{ip},$$

where

$$m_i = n_i^E + n_i^C - 2,$$

and $c(m_i)$ is approximated by

$$c(m_i) = 1 - \frac{3}{4m_i - 1}.$$

1.4 Distribution of Multiple Effect Sizes

For fixed values of δ_{ip} , Hedges (1981) showed that this standardized effect-size estimator, d_{ip}, is asymptotically normally distributed with mean δ_{ip} and variance $\sigma^2(\delta_{ip})$, which can be represented as

$$\sigma^2(\delta_{ip}) = \frac{n_i^E + n_i^C}{n_i^E\, n_i^C} + \frac{\delta_{ip}^2}{2(n_i^E + n_i^C)} .$$

Since δ_{ip} is not known, Hedges (1982a) provided the large sample approximation of $\sigma^2(\delta_{ip})$ by substituting d_{ip} for δ_{ip}. Thus, estimating $\sigma^2(\delta_{ip})$ for the *p*th outcome measure in the *i*th study requires one to replace δ_{ip}^2 by its estimate d_{ip}^2 in the previous equation, or

$$\hat{\sigma}^2(d_{ip}) = Var(d_i|\delta_{ip}) = \frac{n_i^E + n_i^C}{n_i^E\, n_i^C} + \frac{d_{ip}^2}{2(n_i^E + n_i^C)} .$$

Given that this model allows different numbers of effect sizes based on different numbers of outcome measures for each study, the total number of comparisons between experimental and control groups is P, where $P = \Sigma p_i$. As noted above p_i

denotes the number of outcome measures in study i.

Because the measurements for any subject within a study are correlated, the estimated multiple effect sizes will also be correlated. The correlations between the effect sizes, d_{ip}, $p = 1, 2, \ldots, p_i$ in study i, depend upon the correlations between the outcome measures for subjects in the experimental and control groups. However, not all studies report sample correlations among the outcome measures, which force us to impute values for the population correlations from other sources (published test manuals, other studies, etc.). Thus, the covariances between the effect sizes of any two outcome measures p and p' in a study can be calculated using the correlation coefficient between the outcome measures ($\rho_{ipp'}$), the population effect sizes for the pairs of outcome measures, and the sample sizes for the experimental and control groups.

Gleser & Olkin (1994) derived the large sample covariance $\sigma(d_{ip}, d_{ip'})$ between d_{ip} and $d_{ip'}$, which can be calculated as follows

$$\sigma(d_{ip}, d_{ip'}) = \left(\frac{1}{n_i^E} + \frac{1}{n_i^C}\right) \rho_{ip,ip'} + \frac{\frac{1}{2}\, \delta_{ip}\, \delta_{ip'}\, \rho^2_{ip,ip'}}{n_i^E + n_i^C} .$$

Estimating $\sigma(d_{ip}, d_{ip'})$ requires us to replace the effect sizes δ_{ip} by their estimates d_{ip} and to replace $\rho_{ip,ip'}$ by either the calculated sample correlations from each study or the imputed values $r_{ip,ip'}$. Thus,

$$\hat{\sigma}(d_{ip}, d_{ip'}) = \left(\frac{1}{n_i^E} + \frac{1}{n_i^C}\right) r_{ip,ip'} + \frac{\frac{1}{2} d_{ip}\, d_{ip'}\, r_{ip,ip'}^2}{n_i^E + n_i^C} .$$

Thus, having estimated the variances and the covariances of the effect sizes for each study, we obtain the estimated variance-covariance matrix $\hat{\Sigma}_i$ for each study. Its diagonal elements are the variances and the off-diagonal elements are the covariances. By "stacking up" these K covariance matrices along the diagonal of a matrix we get the estimated covariance matrix, $\hat{\Sigma}$, of the sampling errors. So, $\hat{\Sigma}$ is a P by P matrix with $\hat{\Sigma}_i$'s stacked along the diagonal, and all off-diagonal block matrices are zero because we assume that the individual studies are independent. Thus, the matrix $\hat{\Sigma}$ can be represented as

$$\hat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\boldsymbol{\Sigma}}_1 & \underline{0} & \cdot & \cdot & \cdot & \underline{0} \\ \underline{0} & \hat{\boldsymbol{\Sigma}}_2 & \cdot & \cdot & \cdot & \underline{0} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \underline{0} & \underline{0} & \cdot & \cdot & \cdot & \hat{\boldsymbol{\Sigma}}_K \end{bmatrix}.$$

2. PRE-POST MULTIPLE MEASURES FOR EACH STUDY

Another method for estimating effect sizes is using the standardized mean-change measure for pretest-posttest designs outlined by Becker (1988). For multiple outcome measures from each study, the standardized mean-change measure is estimated separately for each of the multiple outcomes for experimental and control samples. For instance, a study with one experimental and one control group for each outcome measure would have two standardized mean changes for each outcome, each computed as the difference in mean performance between the posttest and pretest divided by the pretest standard deviation.

2.1 Estimated Standardized Mean-Change Measure

For each of the K $(i = 1,2,....,K)$ studies, let g_{ip}^{E} and g_{ip}^{C} denote the standardized mean change measures for the experimental and control groups, respectively and can be represented as

$$g_{ip}^{E} = \frac{(\overline{Y}_{ip}^{E} - \overline{X}_{ip}^{E})}{S_{ip}^{E}} \quad \text{and} \quad g_{ip}^{C} = \frac{(\overline{Y}_{ip}^{C} - \overline{X}_{ip}^{C})}{S_{ip}^{C}},$$

where $\overline{X}_{ip}^{E}$ and $\overline{X}_{ip}^{C}$ represent the pretest means for the experimental and control groups, respectively. $\overline{Y}_{ip}^{E}$ and $\overline{Y}_{ip}^{C}$ represent the posttest means for the experimental and control groups respectively. S_{ip}^{E} and S_{ip}^{C} represent their respective pretest standard deviations. For each of the multiple outcome measures, separate standardized mean-change measure were computed for the experimental and control groups.

2.2 Unbiased Standardized Mean-Change Measure

Becker (1988) indicated that these standardized mean change measures are slightly biased estimates of the population standardized mean-change parameters and she derived the unbiased estimates of these standardized mean change measures. The unbiased estimates of the experimental and control standardized mean changes are

$$d_{ip}^{E} = \frac{4(n_{ip}^{E} - 2)}{4n_{ip}^{E} - 5} \left(\frac{\overline{Y}_{ip}^{E} - \overline{X}_{ip}^{E}}{S_{ip}^{E}} \right),$$

and

$$d_{ip}^{C} = \frac{4(n_{ip}^{C} - 2)}{4n_{ip}^{C} - 5} \left(\frac{\overline{Y}_{ip}^{C} - \overline{X}_{ip}^{C}}{S_{ip}^{C}} \right),$$

where n_{ip}^{E} and n_{ip}^{C} are the sample sizes for the experimental and control groups.

2.3 Distribution of Standardized Mean-Change Measure

For fixed values of population standardized mean-change measures, the estimated experimental and control standardized mean-change measures (d_{ip}^{E} and d_{ip}^{C}) are asymptotically normally distributed with mean δ_{ip}^{E} and δ_{ip}^{C} and variances $\sigma^2(\delta_{ip}^{E})$ and $\sigma^2(\delta_{ip}^{C})$, respectively.

Thus, the estimated variances of d_{ip}^{E} and d_{ip}^{C} are

$$Var(d_{ip}^{E}) = \frac{4(1 - r_{XY}^{E}) + (d_{ip}^{E})^2}{2n_{ip}^{E}},$$

and

$$Var(d_{ip}^{C}) = \frac{4(1 - r_{XY}^{C}) + (d_{ip}^{C})^2}{2n_{ip}^{C}}.$$

2.4 Effect Size Estimate

The estimated effect sizes, $\hat{\Delta}_{ip}$, for each outcome measure are the differences between the experimental and control unbiased standardized mean-change measures for each of the outcome measures within each of the K studies and is denoted as

$$\hat{\Delta}_{ip} = d_{ip}^{E} - d_{ip}^{C}.$$

Thus, studies that examine the effects of experimental treatment on p outcome measures will have p effect sizes.

2.5 Distribution of Effect Sizes

For fixed values of Δ_{ip}, the estimate of the asymptotic variance of each of the estimated multiple effect sizes, $\hat{\Delta}_{ip}$, is

$$Var(\hat{\Delta}_{ip}) = \frac{4(1-r_{XY}^{E}) + (d_{ip}^{E})^2}{2n_{ip}^{E}} + \frac{4(1-r_{XY}^{C}) + (d_{ip}^{C})^2}{2n_{ip}^{C}},$$

where r_{XY}^{E} and r_{XY}^{C} are the estimates of the pretest-posttest correlations for the experimental and control groups, respectively.

The covariance between $\hat{\Delta}_{ip}$ and $\hat{\Delta}_{ip'}$ is estimated as

$$Cov(\hat{\Delta}_{ip}, \hat{\Delta}_{ip'}) = r_{ip,ip'} [\sqrt{V(d_{ip}^{E})\ V(d_{ip'}^{E})} + \sqrt{V(d_{ip}^{C})\ V(d_{ip'}^{C})}\,],$$

where $r_{ip,ip'}$ is the estimated correlation coefficient between the pairs of the correlated outcome measures within study i.

As with multiple measures studies, having the estimated variances and the covariances of th effect sizes for each study, we obtain the estimated variance-covariance matrix $\hat{\Sigma}_i$ for each study. Its diagonal elements are the variances and the off-diagonal elements are the covariances. Stacking up these K covariance matrices along the diagonal of a matrix produces the estimated covariance matrix, $\hat{\Sigma}$, of the sampling errors. This $\hat{\Sigma}$ variance-covariance matrix has the same structure as variance-covariance matrix for multiple measures studies developed in the previous section in this chapter.

3. MULTIPLE TREATMENTS FOR EACH STUDY

The model for multivariate mixed meta-analysis for multiple treatment studies assumes that we have K studies each comparing T experimental treatment groups (E_q), $q = 1,2,\ldots.,T$, to a common control group (C). It is important to note that this basic model for multiple treatments can be generalized to situations where we are contrasting T experimental treatment groups without control-group comparisons.

3.1 Population Effect Size

Let the outcome measures Y^E_{ijq} and Y^C_{ij} be normally distributed with means μ^E_{iq} and μ^C_i respectively and a common standard deviation σ^C_i. The corresponding population effect sizes for the treatments within each study are

$$\delta_{iq} = \frac{\mu_{iq}^{E} - \mu_{i}^{C}}{\sigma_{i}^{C}} ,$$

where,

$i = 1,2,\ldots.,K$ studies

and

$q = 1,2,\ldots.,T$ treatment groups.

3.2 Sample Effect Size

The effect sizes δ_{iq} can be estimated by replacing μ_{iq}^{E} and μ_{i}^{C} by their sufficient statistics $\overline{Y}_{iq}^{E}$ and $\overline{Y}_{i}^{C}$ and substituting $\hat{\sigma}_{i}^{C}$ for σ_{i}^{C}. The estimated effect size is

$$g_{iq} = \frac{\overline{Y}_{iq}^{E} - \overline{Y}_{i}^{C}}{\hat{\sigma}_{i}^{C}} ,$$

Here $\hat{\sigma}_i^C$ is the control group standard deviation for study i.

3.3 Distribution of Effect Sizes

For fixed values of δ_{iq} and when the homogeneity of the variances for the multiple treatment groups and the control group holds, the large sample variance of each of the estimated multiple effect sizes d_{iq} (Gleser and Olkin, 1994) is

$$\sigma^2(\delta_{iq}) = \frac{1}{n_{iq}^E} + \frac{1 + \frac{1}{2}\delta_{iq}^2}{n_i^C}.$$

And the population covariances between these correlated multiple effect sizes (Gleser and Olkin, 1994) is

$$\sigma(d_{iq}, d_{iq'}) = \frac{1 + \frac{1}{2}\,\delta_{iq}\,\delta_{iq'}}{n_i^C} .$$

These variances and covariances depend on the effect sizes δ_{iq} and can be estimated by substituting d_{iq} for δ_{iq} (Gleser & Olkin, 1994) and can be calculated as

$$\hat{\sigma}^2(d_{iq}) = Var(d_{iq}|\delta_{iq}) = \frac{1}{n_{iq}^E} + \frac{1 + \frac{1}{2}\,d_{iq}^2}{n_i^C} ,$$

and

$$\hat{\sigma}(d_{iq}, d_{iq'}) = \frac{1 + \frac{1}{2}\,d_{iq}\,d_{iq'}}{n_i^C} .$$

Here we assumed that the variances for the treatment groups and the control group are homogenous. In situations when the homogeneity assumption does not hold, the reader should refer to the article by Gleser and Olkin (1994).

As with multiple measures and pre-post multiple measures studies, having the estimated variances and the covariances of the estimated effect sizes for each study, we end up with the estimated variance-covariance matrix $\hat{\boldsymbol{\Sigma}}_i$ for each study. Its diagonal elements are the variances and the off-diagonal elements are the covariances. Stacking up these K covariance matrices along the diagonal of a matrix produces the estimated covariance matrix, $\hat{\boldsymbol{\Sigma}}$, of the sampling errors. This $\hat{\boldsymbol{\Sigma}}$ variance-covariance matrix has the same structure as the variance-covariance matrix for multiple measures and pre-post multiple measures studies developed in the previous sections in this chapter.

CHAPTER IV

MULTIVARIATE MIXED LINEAR MODEL

As mentioned earlier, Raudenbush and Bryk (1985) developed a univariate empirical Bayes estimation procedure for meta-analysis as an alternative to least squares estimation for the linear model within the formulation of two-stage hierarchical modeling having a prior distribution. Also, Raudenbush, Becker, and Kalaian (1988) developed a multivariate procedure for fixed-effects meta-analysis by using generalized least squares regression to account for the intercorrelations among the multiple outcome measures. Moreover, Raudenbush (1988) reformulated the hierarchical two-stage linear model as a general mixed model where we have missing values in the data set. For instance, not all the studies included in the research synthesis may have the same number of dependent variables or contrasts among treatment groups and the control group because the research interest is different from study to another. In this situation, the studies with missing dependent variables or treatment groups would have to be excluded from the meta-analysis and we would limit the meta-analysis to studies with complete data in order to be able to perform previously developed multivariate statistical meta-anslysis procedures. The alternative analytic method used by reviewers is to use the univariate meta-analysis techniques in order to include and use all the data in the review. As mentioned earlier this practice limits the kind of research questions asked and at the same time inflates Type I error rates.

In this study the univariate empirical Bayes estimation method, the multivariate fixed-effects generalized least squares procedure, and the general mixed model where the data are not of full rank (missing dependent variables in the data set) are combined and extended to the general situation where we have multiple random effect sizes based on multiple dependent variables, multiple correlation coefficients, or multiple treatment groups within each study. The empirical Bayes estimation method will be used to estimate the parameters of the model. The multivariate mixed-effects model is viewed as a two-stage model. At the first stage, the "within-study model" for each individual study having multiple effect sizes is formulated. At the second stage, the parameters of the within-study model are viewed as varying randomly across different studies and some of this variation is thought to be explainable by known study characteristics.

1. WITHIN-STUDY MODEL:

In the within-study model for multivariate mixed meta-analysis, we assume that the observed vector of multiple effect sizes, $\underline{d}_i$, of study i, is equivalent to a vector of population effect sizes $\underline{\delta}_i$ plus a vector of errors, $\underline{e}_i$, for each study. Here the within study variances and covariances are assumed to vary from study to study. Thus, the basic within study model for study i can be represented as

$$\underline{d}_i = X_i \underline{\delta}_i + \underline{e}_i \ , \qquad i = 1,2,\ldots.,K,$$

where,

$\underline{d}_i$ is a vector with $(p_i x1)$ elements,

$\underline{\delta}_i$ is a vector of $(mx1)$ elements,

$\underline{e}_i$ is a vector of $(p_i x1)$ elements,

and,

X_i is a matrix of $(p_i \ x \ m)$ elements of response indicators for the elements of $\underset{\sim}{d}_i$ with $X_i = 1$ when d_i is observed and $X_i = 0$ when d_i is missing.

Here, m $(p_i \leq m)$ is the maximum number of outcome measures across studies assuming that there are no missing effect sizes in the data.

1.1 Illustrative Example

To illustrate the within-study model with missing effect sizes, suppose a reviewer has K studies and most of these studies have two outcome measures. However, some of these studies have only one of the outcome measures and not the other. Thus, this reviewer is faced with the problem of missing data in the research synthesis, especially when trying to build statistical models to explain the variation in these effect sizes. The within-study model for a hypothetical set of K studies can be expressed as follows

$$study\ 1: \quad \begin{bmatrix} d_{11} \\ d_{12} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \end{bmatrix},$$

$$study\ 2: \quad [d_{21}] = [1 \quad 0] \begin{bmatrix} \delta_{21} \\ \delta_{22} \end{bmatrix} + [e_{21}] \ ,$$

$$study\ 3: \quad [d_{32}] = [0 \quad 1] \begin{bmatrix} \delta_{31} \\ \delta_{32} \end{bmatrix} + [e_{32}] \ ,$$

$$\begin{matrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{matrix}$$

$$study\ K: \quad \begin{bmatrix} d_{K1} \\ d_{K2} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \delta_{K1} \\ \delta_{K2} \end{bmatrix} + \begin{bmatrix} e_{K1} \\ e_{K2} \end{bmatrix} .$$

where for example

$$\begin{bmatrix} e_{11} \\ e_{12} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} , \begin{bmatrix} \sigma_{11}^2 & \sigma_{112} \\ \sigma_{121} & \sigma_{12}^2 \end{bmatrix} \right).$$

But the distribution of errors for the second study is

$$e_{21} \sim N(0, \sigma_{21}^2),$$

and for the third study it is

$$e_{32} \sim N(0, \sigma_{32}^2).$$

Stacking the $\underset{\sim}{d}_i$ vectors for all the K studies, produces a single vector, $\underset{\sim}{d}$, containing all the effect sizes. The complete within-study model can be represented as

$$\begin{bmatrix} \underset{\sim}{d}_1 \\ \underset{\sim}{d}_2 \\ . \\ . \\ . \\ \underset{\sim}{d}_K \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \underset{\sim}{X}_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & . & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & . & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & . & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \underset{\sim}{X}_K \end{bmatrix} \begin{bmatrix} \underset{\sim}{\delta}_1 \\ \underset{\sim}{\delta}_2 \\ . \\ . \\ . \\ \underset{\sim}{\delta}_K \end{bmatrix} + \begin{bmatrix} \underset{\sim}{e}_1 \\ \underset{\sim}{e}_2 \\ . \\ . \\ . \\ \underset{\sim}{e}_K \end{bmatrix}$$

which in turn can be expressed in more compact and unsubscripted matrix form as

$$\boldsymbol{d} = \boldsymbol{X\delta} + \boldsymbol{e}\ ,$$

where

$$\boldsymbol{e} \sim \boldsymbol{N}(0\ ,\ \Sigma)\ .$$

Here d is a $Px1$ vector where $P=\sum p_i$, δ is a $Km \times 1$ vector, X is $P \times Km$ matrix of 1's and 0's, and e is a PxP matrix. We further assume that the errors, $\underline{e}_i$, are P-variate normally distributed with a zero mean vector and variance-covariance matrix Σ_i. Thus, Σ is the sampling variance-covariance matrix of the effect sizes for the multiple dependent variables or multiple treatment groups. The formulas for estimating this sampling covariance matrix are shown in Chapter 3 (sections 1.4, 2.4, and 3.3).

1.2 GLS Within-Study Model

The basic within-study model given above can be reformulated as a generalized least squares within-study model (Raudenbush, Becker, and Kalaian, 1988) by factorizing the estimated variance-covariance matrix for each study, $\hat{\Sigma}_i$ which is a positive definite matrix, into the product of a triangular matrix and its transpose (Finn, 1974). This Cholesky decomposition can be represented as

$$\hat{\Sigma}_i = E_i E'_i \ .$$

As an illustration, suppose the variance-covariance matrix for study i, for the bivariate example given in section 1.2, is

$$\hat{\Sigma}_i = \begin{bmatrix} \hat{\sigma}^2(d_{i1}) & \hat{\sigma}(d_{i1}, d_{i2}) \\ \hat{\sigma}(d_{i2}, d_{i1}) & \hat{\sigma}^2(d_{i2}) \end{bmatrix} .$$

Here $\hat{\sigma}(d_{i1}, d_{i2}) = \hat{\sigma}(d_{i2}, d_{i1})$ is the covariance of the two effect sizes in study i, while $\hat{\sigma}^2(d_{i1})$ and $\hat{\sigma}^2(d_{i2})$ are the respective variances of the two effect sizes.

The resulting Cholesky or triangular factor for the i<u>th</u> study is

$$\underline{F_i} = \begin{bmatrix} \hat{\sigma}(d_{i1}) & 0 \\ \hat{\sigma}(d_{i2}, d_{i1}) / \hat{\sigma}(d_{i1}) & \sqrt{\hat{\sigma}^2(d_{i2}) - (\hat{\sigma}^2(d_{i1}, d_{i2}) / \hat{\sigma}^2(d_{i1})} \end{bmatrix} .$$

F_{i22} can be recognized as the conditional standard deviation of d_{i2} given d_{i1}. Thus, the conditional standard deviations, holding the other effect sizes constant, are the diagonal elements of the Cholesky factor matrix. The off-diagonal elements are the conditional covariances, given the other variables or effect sizes.

Premultiplying the within-study model for each study by $\underline{F}_i^{-1}$ yields a set of uncorrelated multiple effect sizes for each study. This can be written as

$$\underline{F}_i^{-1}\underline{d}_i = \underline{F}_i^{-1}\underline{X}_i\underline{\boldsymbol{\delta}}_i + \underline{F}_i^{-1}\underline{e}_i,$$

which in turn can be represented as

$$\underline{d}^*_i = \underline{X}^*_i\underline{\boldsymbol{\delta}}_i + \underline{e}^*_i \ ,$$

where,

$$\underline{e}^*_i \sim N(0 \ , \underline{I}_i) .$$

Here e_i^*, the error vector of the transformed uncorrelated multiple effect sizes for each study, is asymptoticly normally distributed (Hedges, 1981) with mean 0 and identity variance-covariance matrix $\underline{I}_i$. $\underline{I}_i$ consists of 1's in the diagonal and 0's in the off-diagonal.

This within-study model can be rewritten in more general and unsubscripted matrix form as

$$\boldsymbol{d}^* = \boldsymbol{X}^*\boldsymbol{\delta} + \boldsymbol{e}^* ,$$

where

$$e^{*} \sim N(0, I) .$$

Here the matrix I is a block diagonal matrix with identity submatrices in the diagonal.

2. BETWEEN-STUDIES MODEL:

At the second stage, the between-studies model can be formulated into two forms. The first is the unconditional model, where we assume that the multiple effect-size parameters $\underline{\delta}_i$ vary around a grand mean vector plus error. The second is the conditional model, where we assume that the multiple effect-size parameters $\underline{\delta}_i$ depend on known study characteristics plus error.

2.1 Unconditional Between-Studies Model

In this simple basic model, the multiple effect-size

parameters $\underline{\delta}_i$ vary as a function of a grand mean vector (one element for each outcome measure or each treatment group) and random error. The between-studies model for each study can be represented as

$$\underline{\delta}_i = \underline{\Delta} + \underline{U}_i \, , \qquad \underline{U}_i \sim N(\underline{0} \, , \underline{\tau}) \, ,$$

where $\underline{\delta}_i$ and $\underline{U}_i$ are $(m \times 1)$ vectors, and $\underline{\Delta}$ is a vector of grand mean parameters.

The multivariate between-study model for the illustrative example (given in section 1.2) with maximum of $m=2$ outcome measures for each study can be represented as

$$study\ 1: \quad \begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \end{bmatrix} + \begin{bmatrix} U_{11} \\ U_{12} \end{bmatrix} ,$$

$$study\ 2: \quad \begin{bmatrix} \delta_{21} \\ \delta_{22} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{02} \end{bmatrix} + \begin{bmatrix} U_{21} \\ U_{22} \end{bmatrix} ,$$

$$\textit{study } 3\text{:}\quad \begin{bmatrix}\delta_{31}\\ \delta_{32}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\gamma_{01}\\ \gamma_{02}\end{bmatrix} + \begin{bmatrix}U_{31}\\ U_{32}\end{bmatrix},$$

$$\begin{matrix}\cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\\ \cdot & \cdot & \cdot & \cdot\end{matrix}$$

$$\textit{study } K\text{:}\quad \begin{bmatrix}\delta_{K1}\\ \delta_{K2}\end{bmatrix} = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\begin{bmatrix}\gamma_{01}\\ \gamma_{02}\end{bmatrix} + \begin{bmatrix}U_{K1}\\ U_{K2}\end{bmatrix},$$

where,

$$\begin{bmatrix}U_{i1}\\ U_{i2}\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\tau_{i1}^2 & \tau_{i12}\\ \tau_{i21} & \tau_{i2}^2\end{bmatrix}\right).$$

By stacking all the between-study vectors from all the K studies we will have the complete between-study model for all the studies which should be included in the meta analysis. This complete model can be written as

$$\begin{bmatrix} \underline{\boldsymbol{\delta}}_1 \\ \underline{\boldsymbol{\delta}}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\boldsymbol{\delta}}_K \end{bmatrix} = \begin{bmatrix} \underline{I} \\ \underline{I} \\ \cdot \\ \cdot \\ \cdot \\ \underline{I} \end{bmatrix} \boldsymbol{\Delta} + \begin{bmatrix} \underline{U}_1 \\ \underline{U}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{U}_K \end{bmatrix} .$$

This can be rewritten in more general form as

$$\boldsymbol{\delta} = A \, \boldsymbol{\Delta} + U,$$

where

$$U \sim N(0 , \mathbf{T}) , \qquad \mathbf{T} = I_K \otimes \tau .$$

and

$$A = 1_K \otimes I_m.$$

This multivariate linear model allows a different number of outcome measures for each study. When by experimental design not all the m outcome variables are measured in each study, we still can obtain efficient empirical Bayes estimates for the parameters in the model as well as imputed values of the missing d's.

2.2 Conditional Between-Studies Model

In this model, which can be considered an expansion of the unconditional model, we use information about study characteristics (study contexts, study design, treatments, and subject characteristics from each study) to account for the variation among the effect sizes. In other words, we try to explain the variations in the effect-size parameters by knowing methodological and contextual variations in the primary studies in the review under consideration. This

between-study model can be written in the following form

$$\underline{\delta}_i = \underline{W}_i \, \underline{\gamma} + \underline{U}_i , \qquad \underline{U}_i \sim N(\underline{0} , \underline{\tau}) ,$$

where $\underline{\delta}_i$ and $\underline{U}_i$ are vectors having $m \times K$ (m=*maximum* p_i) elements, $\underline{W}_i$ is a $m \times q$ matrix of known study characteristics and $\underline{\gamma}$ is a $q \times 1$ vector of between-studies parameters. Here, we assume that $\underline{U}_i$ has a multivariate normal distribution with mean vector $\underline{0}$ and covariance matrix $\underline{\tau}$.

To illustrate this model we use the illustrative example given in section 1.2 where we have two outcome variables for each study. Here we hypothesize that the two outcome variables in the example are the SAT-Verbal and SAT-Math effect sizes from SAT coaching studies (these SAT coaching studies and the coaching effect sizes are described in detail in Chapter VII). For illustrative purposes, we further hypothesize that the amount of coaching time in hours influences the size of the coaching effect. Thus at the second stage in this conditional model we incorporate information about the number of coaching hours for SAT-Math and SAT-Verbal subtests to explain the variation among the effect sizes from the various SAT coaching studies. This

conditional multivariate between-studies model may be written as

$$\textit{study } 1: \quad \begin{bmatrix} \delta_{11} \\ \delta_{12} \end{bmatrix} = \begin{bmatrix} 1 & W_{11} & 0 & 0 \\ 0 & 0 & 1 & W_{12} \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \\ \gamma_{02} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} U_{11} \\ U_{12} \end{bmatrix},$$

$$\textit{study } 2: \quad \begin{bmatrix} \delta_{21} \\ \delta_{22} \end{bmatrix} = \begin{bmatrix} 1 & W_{21} & 0 & 0 \\ 0 & 0 & 1 & W_{22} \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \\ \gamma_{02} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} U_{21} \\ U_{22} \end{bmatrix},$$

$$\textit{study } 3: \quad \begin{bmatrix} \delta_{31} \\ \delta_{32} \end{bmatrix} = \begin{bmatrix} 1 & W_{31} & 0 & 0 \\ 0 & 0 & 1 & W_{32} \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \\ \gamma_{02} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} U_{31} \\ U_{32} \end{bmatrix},$$

$$\begin{matrix} \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \end{matrix}$$

$$study\ K: \quad \begin{bmatrix} \delta_{K1} \\ \delta_{K2} \end{bmatrix} = \begin{bmatrix} 1 & W_{K1} & 0 & 0 \\ 0 & 0 & 1 & W_{K2} \end{bmatrix} \begin{bmatrix} \gamma_{01} \\ \gamma_{11} \\ \gamma_{02} \\ \gamma_{22} \end{bmatrix} + \begin{bmatrix} U_{K1} \\ U_{K2} \end{bmatrix},$$

where

$$\begin{bmatrix} U_{i1} \\ U_{i2} \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau^2_{i1} & \tau_{i12} \\ \tau_{i21} & \tau^2_{i2} \end{bmatrix} \right).$$

Here W_{i1} and W_{i2} represent hours of coaching in study i

for the SAT-Verbal and SAT-Math outcome measures, respectively.

By stacking all the $\underline{\boldsymbol{\delta}}_i$ vectors for all the studies we will have the complete conditional between-studies model which can be represented in the form

$$\begin{bmatrix} \underline{\boldsymbol{\delta}}_1 \\ \underline{\boldsymbol{\delta}}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{\boldsymbol{\delta}}_K \end{bmatrix} = \begin{bmatrix} \underline{W}_1 \\ \underline{W}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{W}_K \end{bmatrix} \underline{\boldsymbol{\gamma}} + \begin{bmatrix} \underline{U}_1 \\ \underline{U}_2 \\ \cdot \\ \cdot \\ \cdot \\ \underline{U}_K \end{bmatrix} .$$

This can be rewritten in more general and unsubscripted matrix form as

$$\boldsymbol{\delta} = \boldsymbol{W}\boldsymbol{\gamma} + \boldsymbol{U} \quad ,$$

where

$$\boldsymbol{U} \sim N(0\ ,\ \mathrm{T})\ .$$

Therefore, $\mathrm{T} = I_K \otimes \tau$, is the conditional covariance matrix of the multiple effect sizes. In other words, it is the amount of unexplained parameter variation and covariation left after knowing the effects of coaching hours. From the Bayesian point of view, this second stage model is considered the prior distribution of δ.

3. WITHIN-STUDY AND BETWEEN-STUDIES MODELS COMBINED

Combining the within and between study models for each individual study, we get

$$\underline{d}_i = \underline{X}_i \underline{W}_i \underline{\gamma} + \underline{X}_i \underline{U}_i + \underline{e}_i \ ,$$

The transformed combined model for each study can be written as

$$d^*_i = X^*_i W_i \gamma + X^*_i U_i + e^*_i ,$$

which in turn can be rewritten in more general matrix form as

$$d^* = X^* W \gamma + X^* U + E^* .$$

In this combined model we assume that U and E^* are independent and can be considered as a specific case of the general mixed linear model which can be represented as

$$d^* = A_1 \theta_1 + A_2 \theta_2 + R ,$$

where d^* is a vector of the uncorrelated multiple effect sizes, θ_1 is a vector of unknown fixed effects parameters, θ_2 is a vector of unknown random effects parameters, A_1 and A_2 are known matrices of study characteristics, and R is a block diagonal matrix of error terms.

In this mixed linear model, the Bayesian view is to assume that the fixed effects parameters as having prior distribution that is normal with zero mean vector and variance-covariance matrix Γ. Also, Γ is assumed to be infinitely large. Thus, Γ^{-1} is close to 0. That is

$$\theta_1 \sim N(0,\Gamma) .$$

Further, we assume that the vector of the random effects parameters are normally distributed with zero mean vector and variance-covariance matrix $\mathbf{T}$. That is

$$\theta_2 \sim N(0,\mathbf{T}) ,$$

and

$$R \sim N(0, \Sigma) \ .$$

Also, we assume that the parameters θ_1, θ_2, and R are mutually independent vectors.

Comparing the two stage linear model and the general mixed linear model we can see that

$$A_1 = X^* W \quad , \quad A_2 = X^* \ ,$$

$$\theta_1 = \gamma \quad , \quad \theta_2 = U \ .$$

Consequently these two models can be considered as a special case of a single general Bayesian linear model which can be written as

$$\boldsymbol{d}^{*} = A\,\theta + E^{*} \;, \qquad E^{*} \sim N(0\;,\;\Psi) \;,$$

where d^{*} is the outcome vector, A is the predictor matrix, θ is the parameter vector, and we assume that the prior distribution of θ is

$$\theta \sim N(\bar{\theta}\;,\;\Omega)\;,$$

where

$$\Omega = \begin{bmatrix} \Gamma^{-1} & 0 \\ 0 & \mathbf{T} \end{bmatrix} .$$

Now comparing this general linear model with the mixed linear model and the two stage linear model we can see that

$$A = [\, A_1 \mid A_2 \,] \quad , \quad \theta = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \quad ,$$

or

$$A = [\, X^* W \mid X^*] \quad , \quad \theta = \begin{bmatrix} \gamma \\ U \end{bmatrix} \quad .$$

CHAPTER V

ESTIMATION OF THE MULTIVARIATE MIXED MODEL

This chapter provides a description of the estimation of the parameters of the multivariate mixed-effects model for effect-size meta-analysis and research reviews. Also, in this chapter the posterior means and variance-covariance matrices for the parameters in the model are presented.

In the second section of this chapter, the maximum likelihood estimates of the dispersion matrices for multivariate effect-size data are presented. These maximum likelihood estimates are obtained using the EM algorithm.

1. ESTIMATION WHEN T AND Σ ARE KNOWN

The two-stage multivariate linear model for effect sizes developed in the previous chapter is

Within-study model:

$$d^* = X^*\delta + e^*, \qquad e^* \sim N(0,I) \ .$$

Between-studies model:

$$\delta = W\gamma + U \ , \qquad U \sim N(0,\mathrm{T}) \ .$$

Combined-model:

$$d^* = X^*W\gamma + X^*U + e^* \ .$$

Here,

$$var(d^*) = X^* \ \mathrm{T} \ X^{*\prime} + I \ ,$$

and

$$[var(d^*)]^{-1} = I - X^*(X^{*\prime}X^* + T^{-1})^{-1}X^{*\prime} .$$

Given the above two-stage multivariate linear model, the Bayesian point of view considers the second stage to be the prior distribution for δ and adds another stage which specifies the prior distribution of γ as being normal with mean vector 0 and variance-covariance matrix Γ. We further assume that Γ is infinitely large. Thus, Γ^{-1} is approximately 0.

1.1 Posterior Distribution of $\theta=(\gamma,U)'$

Given the above assumptions and the fact that the data are not of full rank (because of missing outcome variables), the formulas derived by Raudenbush (1988) apply here. Thus, the posterior distribution of $\theta=(\gamma,U)'$ given d, T, I is normal with posterior mean vector

$$\theta^* = \begin{bmatrix} \gamma^* \\ U^* \end{bmatrix},$$

and posterior dispersion (variance-covariance) matrix

$$D_\theta^* = \begin{bmatrix} D_\gamma^* & C_{\gamma U}^* \\ C_{\gamma U}^* & D_U^* \end{bmatrix},$$

where,

$$D_\gamma^* = D(\gamma \mid d^*, \mathrm{T}) = [W'X^{*\prime}X^*W - W'X^{*\prime}X^*(X^{*\prime}X^* + \mathrm{T}^{-1})^{-1}X^{*\prime}X^*W]^{-1}$$

$$=[W'(X^{*\prime}X^* - X^{*\prime}X^*(X^{*\prime}X^* + \mathrm{T}^{-1})^{-1}X^{*\prime}X^*)W]^{-1} ,$$

$$= [W'(I-\Lambda')X^{*\prime}X^*W]^{-1}$$

$$= [W' \Lambda T^{-1} W]^{-1}$$

here,

$$\Lambda = (X^{*\prime} X^{*} + T^{-1})^{-1} X^{*\prime} X^{*} .$$

$$D_U^{*} = D(U \mid d^{*}, T) = C^{-1} + C^{-1} X^{*\prime} X^{*} W D_{\gamma}^{*} W' X^{*\prime} X^{*} C^{-1}$$

$$= C^{-1} + \Lambda W D_{\gamma}^{*} W' \Lambda' . ,$$

where,

$$C^{-1} = (X^{*\prime} X^{*} + T^{-1})^{-1} .$$

$$C_{\gamma U}^{*} = -D_{\gamma}^{*} W' X^{*\prime} X^{*} C^{-1}$$

$$= -D_{\gamma}^{*} W' \Lambda' \, .,$$

and

$$C_{U\gamma}^{*} \quad = \quad (C_{\gamma U}^{*})' \quad .$$

Now after finding the estimates of the posterior variance-covariance matrix we can find the estimates of the posterior expectations of γ and U which are

$$\gamma^{*} \quad = \quad E(\gamma \mid d^{*}, \mathrm{T}) \quad = \quad D_{\gamma}^{*}[W' X^{*\prime} d^{*} \; - \; W' X^{*\prime} X^{*} C^{-1} X^{*\prime} d^{*}]$$

$$= \quad D_{\gamma}^{*} W' (I - X^{*\prime} X^{*} C^{-1}) X^{*\prime} d^{*}$$

$$= \quad D_{\gamma}^{*} W' \; (I - \Lambda') \; X^{*\prime} d^{*}$$

$$U^{*} \quad = \quad E(U \mid d^{*},\mathrm{T}) \quad = \quad C^{-1}X^{*\prime}(d^{*} - X^{*}W\gamma^{*}) \ .$$

1.2 Posterior Distribution of δ

From the above estimates, the posterior expectation of δ is

$$\delta^{*} \quad = \quad E(\delta \mid d,\mathrm{T}) \quad = \quad E[(W\gamma \ + \ U) \mid d^{*},\mathrm{T}]$$

$$= \quad X^{*}W\gamma^{*} \ + \ X^{*}C^{-1}X^{*\prime}(d^{*} - X^{*}W\gamma^{*})$$

Last, the posterior dispersion matrix of δ^{*} is

$$D_{\delta^*} = D(\delta|d^*,\mathrm{T}) = D(W\gamma|d^*,\mathrm{T}) + D(U|d^*,\mathrm{T}) + cov(W\gamma,U|d^*) + cov(U^*,W\gamma$$

$$= C^{-1} + (I-\Lambda)X^*D_\gamma^*W'X^{*'}(I-\Lambda)' .$$

2. M.L.E. ESTIMATION OF THE DISPERSION MATRICES VIA EM

The empirical Bayes estimation procedure discussed above assumed that the covariance matrices Σ and T are known. But in real situations this is not the case. Consequently δ_i and γ cannot be estimated using the formulas above because their maximum likelihood estimates do not exist in closed form especially in the unbalanced case and data sets with missing data points (Dempster, Laird, & Rubin, 1977). Thus, from the empirical Bayes point of view, point estimates of the dispersion parameters are first calculated. Then these point estimates are substituted into the formulas for calculating the posterior expectation and dispersion matrices. Typically, these dispersions are estimated by means of maximum likelihood, so that they will be asymptotically efficient with known large sample normal distributions as $K \to \infty$ (Raudenbush, 1988).

Dempster, Laird, and Rubin (1977) and Little and Rubin (1987) suggested the use of EM algorithm as a numerical approach to compute maximum likelihood point estimates of the unknown variance and covariance components from incomplete data. Pigott (1992, 1994) outlined the EM algorithm procedure to obtain the maximum likelihood estimates for effect-size data with missing predictor data points.

In this study we consider the case of the multivariate data to be incomplete and missing (ignorable nonresponse) because individuals in the primary research studies are observed on different subsets of the complete set of variables (Little and Rubin, 1987). Here, the EM algorithm developed by Dempster, Laird, and Rubin is used to obtain the maximum likelihood estimates of the parameters. Using this method of estimation, we assume that the population effect size vector δ is known. Further, we assume that the errors of the between-study model, U, have been observed. Given these assumptions, the covariance matrix T can be estimated by

$$\hat{\mathrm{T}} = K^{-1} \sum U_i U_i' .$$

The basic idea of the EM algorithm is to estimate the "complete data" sufficient statistics ($\sum U_i U_i'$) and then find the maximum likelihood estimates of T based on the estimates of the sufficient statistics. Of course, estimation of the sufficient statistics requires initial estimates of the covariance matrix, T by using, for example, ordinary least squares estimates of the residuals from within-study and between- studies models. Thus the goal of the EM procedures is to find parameter estimates based on the expected values for the sufficient statistics of the statistical model.

The EM algorithm is an iterative procedure where each iteration consists of two steps (estimation and maximization steps). I next illustrate how the two step process of this algorithm works.

2.1 E-Step (Expectation Step)

Given the initial estimates of T and the effect size estimates, we can find the posterior expectation of the sufficient statistic $\sum U_i U_i'$ of the model as

$$E(\sum U_i U_i^{\prime} \mid d^*, \mathrm{T}) = \sum U_i^{*\prime} U_i^* + \sum (X_i X_i^{\prime} + \mathrm{T}^{-1})^{-1} + \sum \Lambda_i W_i D_\gamma^* W_i \Lambda_i^{\prime} .$$

Here T refers to the initial estimate of T, we also assume that the multiple effect sizes have a multivariate normal distribution. This multivariate normal distribution has sufficient statistics which are the sums and the sums of the crossproducts of the observations in the data. Conditional expectations of these sufficient statistics are used to estimate the mean vector and the variance-covariance matrix of the multivariate normal distribution.

2.2 M-Step (Maximization Step)

Based on the expected values of the sufficient statistics from the E-step, new estimates of the elements of the covariance matrix T are computed. At the end of the iterative process (estimation and maximization steps) the estimate of the matrix T converges to local maximum (Dempster, Laird, and Rubin, 1977; Little and Rubin, 1987; Pigott, 1994). This new

T matrix estimate can be substituted in the formulas for finding the posterior mean vector and variance-covariance matrices of δ, γ, and U.

In summary, the E-step of the EM algorithm produces the posterior expectations of the complete data sufficient statistics at each stage of the iteration. This expected value of the sufficient statistic ($E(\sum U_i^{\prime} U_i)$) can be used to find new estimates of T.

Once this new covariance matrix is found, it can be substituted in the formulas for finding the posterior mean vectors and dispersion matrices of δ, γ, and U. Then, these new posterior values can be substituted in the formula for finding the expected value of the sufficient statistics to yield a new posterior expectation (E-step). This new posterior expectation produces a new posterior estimate of T (M-step). The resulting value of T is then used as input for the next E-step. The process iterates back and forth until convergence to the maximum likelihood estimates at a required degree of accuracy is attained.

CHAPTER VI

EMPIRICAL APPLICATION OF MULTIVARIATE HIERARCHICAL LINEAR MODEL

The proposed multivariate hierarchical linear model and estimation theory are applied in this chapter. The study involves the analysis of generated multivariate normal data set with pre-specified parameter values using FORTRAN program and IMSL (Version 10) subroutines.

The main rationale for using an artificial data set in this study was to validate the multivariate estimation procedures using the HLM computer program for analyzing multi-level data sets versus using the V-known program which is part of the HLM program. The V-Known program was designed for analyzing effect-size sets for research-synthesis purposes and it can be used for univariate (one effect-size from each study) and complete multivariate (multiple correlated effect sizes from each study with no missing effect sizes) meta-analyses. In other words, the existing V-known program can be used for multivariate meta-analysis when we have the same number of multiple correlated effect sizes from each study in the review. However, this is typically not the case.

1. INTRODUCTION TO THE HLM COMPUTER PROGRAM

The hierarchical linear model (HLM) program (Bryk et al., 1986) applies the EM algorithm to provide restricted maximum likelihood (RML) estimates of the variance-covariance components (Dempster, Laird, and Rubin, 1977). Consequently, these estimates of the variances and the covariances can be used to obtain empirical Bayes estimates of the linear model parameters.

This program (Bryk et al., 1986) is available to researchers from different disciplines. It constitutes a general analytic method for studying multi-level data with hierarchical characterization and analyzing effect-size data sets for meta-analysis and research review purposes.

Using the HLM program for meta-analysis and research synthesis typically involves the application of the V-Known routine (Bryk et al., 1986) within the HLM computer program. The V-Known routine is a general multivariate regression routine for univariate and multivariate effect-size data sets(data with the same number of effect sizes for each study) and assumes that the sampling variance-covariance matrix among these multiple effect-size parameters is known. However, the V-Known routine for analyzing research-synthesis data cannot handle multivariate data sets with missing effect sizes for some of the studies and complete data for the rest of the studies.

In this study, the HLM program for analyzing multi-level hierarchical data is used to estimate the parameters of the multivariate mixed-effects model for meta-analysis with missing data points. Before applying the HLM program, the within-study model is reformulated as a weighted least-squares within-study model by using the Cholesky factorization principle.

The procedure for reformulating the within-study model involves the following steps. First, each of the estimated variance-covariance matrices for the vector of estimated multiple effect sizes from each of the primary studies in the review should be factorized to a Cholesky triangular matrix. Second, the components of the within-study model (in this case the vectors of the multiple effect sizes and the identity matrices) are premultiplied by the inverse of the resulting Cholesky triangular factors. After these two reformulation steps, the HLM program for multi-level data (not the V-Known program) can be used to fit the specified statistical model and obtain empirical Bayes estimates of the parameters in the model.

2. MULTIVARIATE EFFECT-SIZE DATA GENERATION

The proposed multivariate mixed-effects model which was presented in previous chapters allows for K studies with p_i effect sizes from each study. In other words, it allows different numbers of multiple effect sizes and different number of predictors from each study.

Previous research in meta-analysis suggested values and ranges for the parameters of the generated data in this study.

A. <u>Number of Studies</u>

In fact, not many published meta-analysis deal with multivariate effect sizes because of the complexity of the data and the statistical analysis for such data. The reviews of the SAT coaching studies (Becker, 1990; Kalaian & Raudenbush, 1994) reported 20 studies and 47 samples with SAT-Verbal and SAT-Math effect sizes. Based on these reviews, the number of studies chosen for this simulation study was K = 50.

B. Sample Sizes for each Study

Hedges and Olkin (1985) used for their simulation studies sample size values ranging from 10 to 100 for each of the experimental and control groups in each of the ***K*** studies. On the other hand, the SAT coaching studies reviewed previously (Becker, 1990; Kalaian & Raudenbush, 1994) contained eight studies with sample sizes larger than 100 for the coached and uncoached groups.

Based on these findings, 50 sample sizes ranging from 10 to 150 were generated from a uniform distribution (10 , 150).

C. Multiple Effect Sizes

For simplicity of interpreting the results from applying the hierarchical linear model to multivariate effect sizes, a data set with bivariate effect sizes was chosen in this data generation. However, the procedure and data analysis can be generalized to data sets with more than two effect sizes in each study.

The artificial bivariate effect sizes (two outcome variables or effect sizes from each study) for this study was generated from the multivariate normal distribution with zero mean vector and the following variance-covariance matrix

$$\Sigma_i = \begin{bmatrix} 0.0800 & 0.0528 \\ 0.0528 & 0.0800 \end{bmatrix} .$$

The values of these variances were chosen based on: (a) the results of previous simulation research in meta-analysis. For instance, Hedges and Olkin (1985) reported the variance of effect sizes when the sample size is 100 as being equal to 0.083 (TABLE 3, P.84) and (b) the results of synthesizing SAT coaching effectiveness studies. For example, Kalaian and Raudenbush (1994) synthesized SAT coaching data set, where the average of the variances of the effect sizes was about 0.07. This chosen variance of 0.08 corresponds to a standard deviation of about 0.28. So about 95% of the effect sizes would be between -0.6 and 0.6 if the mean of the effect sizes is zero.

Then, based on these values of the variances, the covariances between the effect sizes were calculated using the formula given in Chapter 3. In calculating these covariances, a value of 0.66 was used as the correlation between the two correlated effect sizes. Again, this value was chosen because the actual correlation between SAT-Verbal and SAT-Math scores is 0.66.

3. RESULTS

In order to validate the workability and the appropriateness of using the HLM computer program instead of the V-Known program for research synthesis purposes, the V-known and the HLM program were applied to the artificial generated bivariate data. These artificial bivariate effect sizes, sample sizes for the experimental and control groups, and the estimated variances and covariances of these effect sizes are listed in Table 2.

3.1 DESCRIPTION OF THE GENERATED DATA

The values of the set of generated effect sizes for the 50 samples (Table 2) ranged from -0.71 to 0.65 in standard deviation units with an overall average of 0.03 and a variance of 0.07, while the second set of effect sizes ranged from 0.63 to 0.47 with an overall average of -0.04 and variance 0.068. Thus, the average effect sizes for both generated effect-size sets appear to be quite similar and equal to the elements of the population mean vector of the bivariate normal distribution (which was set to zero).

3.2 THE V-KNOWN PROGRAM RESULTS

APPENDIX A contains the listing of the computer output from applying the V-Known program to these generated artificial data. The results of fitting the unconditional regression model to the generated multiple effect sizes using the V-Known program show that the average of the first set of effect sizes (labeled V in Table 6.1) is 0.04 (p = 0.276) and it is -0.02 (p = 0.320) for the second set of effect sizes which is labeled M.

Furthermore, the findings show that the estimated variance-covariance matrix of the random effects part of fitting the unconditional hierarchical linear model is

$$\hat{\mathbf{T}} = \begin{bmatrix} 0.05411 & 0.02010 \\ 0.02010 & 0.03614 \end{bmatrix} .$$

Finally, the results show that both of these variance components are significant (p < 0.000). These findings indicate that the V and M effect sizes are inconsistent across the 50 studies.

3.3 THE HLM PROGRAM RESULTS

APPENDIX B documents the computer output of applying the HLM program and the proposed multivariate mixed linear model to these generated data (bivariate effect-size data set). The results of fitting the unconditional hierarchical linear model to these generated data using the HLM program show that the average effects of V is 0.04 (p = 0.276) and it -0.02 (p = 0.320) for M. These results are identical to those from the V-Known routine.

Also, these results indicate that the estimated variance-covariance matrix of the random effects part of fitting the unconditional hierarchical linear model is

$$\hat{\mathbf{T}} = \begin{bmatrix} 0.05411 & 0.02010 \\ 0.02010 & 0.03614 \end{bmatrix} .$$

Finally, the results show that both of these variance components are significantly different from zero ($p < 0.000$). These findings indicate that the V and M effect sizes are inconsistent across the 50 studies.

4. CONCLUSIONS

The results of applying the V-Known and the HLM programs to the same generated artificial data set yielded exactly the same parameter estimates. The fixed and random-effects parameter estimates of the proposed hierarchical linear model in this study were exactly the same.

In summary, we learned from these two applications is that the HLM program, which is designed for analyzing multi-level data sets, can be used for multivariate meta-analysis purposes. Consequently, this HLM program application can be used to analyze multivariate effect-size data sets with missing data points using the mixed-effects model developed in chapter 4.

Table 2

Generated Multivariate Effect Sizes

STUDY	n^E	n^C	V	M	v(V)	cov(V,M)	v(M)
study 1	28	28	-0.17	-0.27	0.07	0.05	0.07
study 2	80	80	-0.71	-0.26	0.03	0.02	0.03
study 3	105	105	0.46	0.35	0.02	0.01	0.02
study 4	110	110	0.30	-0.28	0.02	0.01	0.02
study 5	40	40	-0.34	-0.21	0.05	0.03	0.05
study 6	64	64	-0.23	-0.19	0.03	0.02	0.03
study 7	91	91	-0.01	-0.11	0.02	0.01	0.02
study 8	47	47	-0.17	-0.27	0.04	0.03	0.04
study 9	85	85	0.23	0.47	0.02	0.02	0.02
study 10	52	52	0.20	-0.07	0.04	0.03	0.04
study 11	137	137	-0.14	-0.00	0.01	0.01	0.01
study 12	48	48	0.14	0.26	0.04	0.03	0.04
study 13	70	70	-0.22	-0.19	0.03	0.02	0.03
study 14	47	47	0.38	0.33	0.04	0.03	0.04
study 15	72	72	-0.19	-0.09	0.03	0.02	0.03
study 16	148	148	-0.12	-0.19	0.01	0.01	0.01
study 17	38	38	0.31	-0.16	0.05	0.03	0.05
study 18	47	47	0.17	0.46	0.04	0.03	0.04
study 19	34	34	-0.18	0.43	0.06	0.04	0.06
study 20	52	52	-0.02	0.01	0.04	0.03	0.04
study 21	146	146	0.24	0.09	0.01	0.01	0.01
study 22	46	46	-0.15	-0.63	0.04	0.03	0.05
study 23	12	12	-0.09	-0.59	0.17	0.11	0.17
study 24	128	128	0.43	0.37	0.02	0.01	0.02
study 25	67	67	0.26	-0.19	0.03	0.02	0.03
study 26	140	140	0.10	0.09	0.01	0.01	0.01
study 27	94	94	-0.13	-0.29	0.02	0.01	0.02
study 28	32	32	0.19	0.02	0.06	0.04	0.06
study 29	105	105	0.65	0.31	0.02	0.01	0.02
study 30	91	91	0.22	0.18	0.02	0.01	0.02
study 31	106	106	0.43	0.19	0.02	0.01	0.02
study 32	111	111	0.04	0.31	0.02	0.01	0.02
study 33	22	22	0.02	-0.12	0.09	0.06	0.09
study 34	29	29	-0.04	-0.05	0.07	0.05	0.07
study 35	65	65	-0.22	-0.07	0.03	0.02	0.03
study 36	141	141	-0.05	0.11	0.01	0.01	0.01
study 37	145	145	0.45	-0.16	0.01	0.01	0.01
study 38	50	50	-0.46	-0.02	0.04	0.03	0.04
study 39	72	72	-0.00	0.08	0.03	0.02	0.03
study 40	139	139	0.15	-0.26	0.01	0.01	0.01
study 41	19	19	-0.08	0.09	0.11	0.07	0.11
study 42	14	14	0.29	-0.14	0.14	0.09	0.14
study 43	88	88	-0.17	-0.02	0.02	0.02	0.02
study 44	132	132	-0.40	-0.38	0.02	0.01	0.02
study 45	14	14	0.19	0.00	0.14	0.09	0.14
study 46	137	137	-0.39	-0.30	0.01	0.01	0.01
study 47	143	143	0.18	-0.20	0.01	0.01	0.01
study 48	88	88	0.04	-0.23	0.02	0.01	0.02
study 49	14	14	-0.23	-0.43	0.14	0.10	0.15
study 50	144	144	0.10	0.24	0.01	0.01	0.01

CHAPTER VII

SAT-COACHING EFFECTIVENESS: A META-ANALYSIS USING MULTIVARIATE HIERARCHICAL LINEAR MODEL

SAT coaching studies are used in this chapter to illustrate the application of the multivariate mixed-effects linear model for meta-analysis with missing effect sizes. This model was developed in chapter 4 and tested using generated bivariate effect sizes (chapter 6). The purposes of the present application is (a) to show the applicability of the proposed model to educational research and multivariate meta-analysis with missing data points, and (b) to compare the results and parameter estimates of applying the multivariate mixed-effects linear model to SAT coaching studies with the results of applying the multivariate fixed-effects model to this data set.

1. INTRODUCTION

In 1926 the Scholastic Aptitude Test (SAT) was first introduced into the College Board's admissions testing program (Dyer, 1987). Nearly a thousand of the nation's colleges and universities now require the SAT examination, and each year approximately a million high school students take the SAT as one of their main college admission requirements. As a result of the importance of the SAT for college entrance, some secondary schools have been importuned by students, parents, and school counselors to provide SAT coaching sessions and test-preparation courses. At the same time, commercial coaching schools have promised the public to increase students' SAT scores dramatically within a short period of time through their special coaching programs (Kalaian & Becker, 1986).

Over the last forty years a great deal of controversy has emerged about the effectiveness of coaching for the Scholastic Aptitude Test. The Educational Testing Service (ETS), which has been developing and administering the SAT, claims that coaching and training programs have little effect in raising students' SAT scores. Their argument relies on the fact that aptitude tests measure cognitive and intellectual skills such as quantitative problem solving and verbal reasoning skills which develop gradually over the years as a result of various experiences (in-school, out-of-school, and in the home). Consequently, they say that SAT scores do not depend upon a specific course of study or highly focussed verbal and mathematical content teaching. Commercial coachers, on the other hand, claim that special SAT coaching classes, test preparation manuals, instruction in test-taking strategies, drill and practice on SAT test items, and test familiarization can yield significant increases in the mastery of the cognitive and analytical skills tested by the SAT and consequently increases in a students' SAT scores.

2. DESCRIPTION OF THE SCHOLASTIC APTITUDE TEST (SAT)

The Scholastic Aptitude Test (SAT) is "a multiple-choice test of how well one has acquired the ability to reason expeditiously with the kind of verbal and mathematical facts and concepts one has presumably acquired in elementary and secondary schools" (Dyer, 1987). It consists of an 85-item verbal subtest (SAT-V) and a 60-item mathematics subtest (SAT-M). The verbal subtest measures vocabulary, reading comprehension, and verbal reasoning. On the other hand the mathematics subtest measures mathematical reasoning and comprehension abilities in the areas of arithmetic, algebra, and geometry and problem solving skills (Comras, 1984).

3. PAST RESEARCH ON SAT COACHING EFFECTIVENESS

In the last 13 years, six studies have reviewed and summarized the results from primary SAT coaching effectiveness studies. The first review was by Slack and Porter (1980), who reviewed 10 reports published prior to 1968. They calculated the mean gain scores for SAT-V and SAT-M subtests separately and compared the results for studies which had used either experimental or statistical controls to those studies without comparison groups. The average gain score for the controlled studies was 16 points for the SAT-V and 12 points for the SAT-M. When the results of all the studies (controlled and uncontrolled studies) combined, the average gains were 29 points for the SAT-V and 33 points for the SAT-M. Clearly the uncontrolled studies in their review produced greater gains than did experimentally or statistically controlled studies. Consequently Slack and Porter concluded that coaching can effectively help students to raise their scores and they stated that "there is ample evidence that students can successfully train for the SAT and that the more time students devote to training, the higher their scores will be" (p. 164).

The second review was conducted by Messick and Jungeblut (1981), who included SAT primary coaching studies published prior to 1980, but excluded two SAT-M and two SAT-V studies used by Slack and Porter. They studied the relationship between the number of coaching hours and the size of coaching effects using regression analyses and they concluded that logarithmically transformed student contact hours were linearly related to coaching effects (gain scores). But the slope coefficient for the regression of SAT-M gain scores on logarithmic transformed contact time was steeper than the SAT-V slope coefficient. In their review they also distinguished between controlled and uncontrolled studies. The average gain scores for experimental (coached) groups weighted by the group sample sizes were 14.3 for SAT-V and 15.1 for SAT-M. Contrary to the experimental studies, the average gain scores in all studies, both experimental and non-experimental (weighted by the control-group sample sizes) were 38 points for the SAT-V and 54 points for the SAT-M.

Dersimonian and Laird (1983) conducted a third review in which they incorporated all the studies used by Slack and Porter and Messick and Jungeblut. Their approach differed from those of the two previous reviews because it involved the use of a random-effects model to estimate the effects of coaching and explain the variability in the coaching effects across studies. That is, they separated the true variation in coaching effectiveness from the within-study sampling variation. They reported that uncontrolled studies had gain scores three times larger than controlled studies and five times larger than matched or randomized studies for both SAT-V and SAT-M subtests. Consequently, they concluded that coaching has positive effects on SAT scores, but the size of the coaching effect is too small to be practically important.

The fourth synthesis was by Kulik et al. (1984), who reviewed only the controlled studies (a total of 14 studies). They calculated standardized mean differences (effect sizes) for each study and concluded that coaching raised SAT scores by 0.21 standard deviation units in four randomized studies versus 0.12 standard deviations for non-randomized studies.

Kalaian and Becker (1986) conducted the fifth review in which they utilized multivariate techniques to analyze the SAT coaching studies. Their results showed considerable variability of effect sizes among SAT coaching studies and that duration of coaching and sponsorship by the Educational Testing Service (ETS) predicted effect size. Their multivariate findings indicated that the effect of coaching is to increase SAT-Math scores by about 18 points and SAT-Verbal scores by 17 points.

The sixth and the last review was by Becker (1990), who reviewed 23 coaching effects reports utilizing the standardized mean-change measure for pretest-posttest research together with the generalized least-squares (GLS) approach for modeling multivariate study outcomes (SAT-V and SAT-M). In this review, studies without control groups are included. The results showed stronger coaching effects for the SAT mathematical subtest. Furthermore, regression models based on published research showed nonsignificant residual variance with coached groups exceeding control groups by 0.09 standard deviations on SAT-V and 0.16 on SAT-M.

In summary, although each of the previous reviews examined different sets of studies and used different quantitative methods to summarize the results of the coaching effectiveness studies, they also shared common conclusions. For example: (a) Studies without a control group have higher coaching effects than controlled studies; (b) There is a remarkable amount of variation in outcomes of SAT coaching studies; (c) Duration of coaching intervention is strongly related to coaching effects; (d) There is a differential effect of coaching on SAT-V and SAT-M subtests.

This review considers the controlled primary studies reviewed previously and more recent primary studies using multivariate mixed-effect approach for meta-analysis (Kalaian, 1994). By using this approach I will be able to

1. investigate and model the variation in the multiple outcomes (SAT-V and SAT-M) simultaneously as a function of study, sample, and coaching characteristics;

2. estimate the variance-covariance of the multiple random effects and test the hypothesis of no variation-covariation among the multiple effect size parameters;

3. estimate the relationships between study characteristics and the multiple study outcomes;

4. include in the analysis the primary studies that coached for both SAT subtests (verbal and mathematical) as well as studies that coached for only one of the subtests and not the other; and

5. include in the analysis different predictors for the SAT-V and SAT-M outcomes. For example, different contact coaching hours for SAT-V and SAT-M within a single study.

6. to use the GLS transformed within-study model (proposed in Chapter 4 to perform multivariate fixed-effects statistical analysis.

4. METHODOLOGY

4.1 Studies in the Review

The set of primary studies reviewed here include those studies examined by previous reviews plus new studies retrieved through a search of the Educational Resources Information Center (ERIC) database. However, our analysis uses only the randomized, matched, and statistically controlled studies. The studies without control groups were excluded from this synthesis for two reasons: (1) Previous reviews showed much higher coaching effects for uncontrolled studies than controlled studies; (2) The effect sizes from the uncontrolled studies in previous reviews showed considerable variability and sometimes more than twice the variability in the controlled studies; (3) Such studies lack internal validity. Uncontrolled studies by Coffin-second experiment (1987), Johnson (1984), Coffman & Parry (1967), Marron (1965), Pallone (1961) are excluded. Also, in this review we consider only the results for the subtest for which coaching and instruction was provided because some studies coached and provided instruction for only one SAT subtest but examined the subjects on both subtests (for example, French, 1958).

Furthermore, many primary studies reported results of coaching effects from different schools or used different coaching programs (e.g., Alderman & Powers, 1980; Evans & Pike, 1973). Because different subgroups of students were involved in the comparisons between coached and control groups, we have treated the effect-size estimates calculated from separate schools within each study as distinct and independent samples. As a result, we identified 39 samples in which SAT-Verbal subtest is coached and tested and 28 samples examining SAT-Math. Only 20 from these two groups examined coaching effects for both SAT subtests (Table 3).

4.2 Study Features

Study characteristics may be coded as part of any meta-analysis technique in order to explain the sources of the variations in the effect sizes. Here, these characteristics included experimental design, context, and subject characteristics. Table 4 lists and summarizes the features and the characteristics of the studies considered in this review.

4.3 Statistical Procedures

The pre-post multiple measures procedure for pretest-postttest designs outlined in the second section of Chapter 3 is used in this review to measure the effectiveness of SAT coaching. The standardized mean change measure is computed separately for each of the SAT-V and SAT-M coached and uncoached samples. For instance, a study with one coached and one uncoached group for each SAT subtest (SAT-V and SAT-M) would have two satandardized mean changes for each outcome, each computed as the difference in mean performance between the posttest and pretest divided by the pretest standard deviation.

Let g_i^C and g_i^U denote the standardized mean change measures for coached and uncoached groups respectively for each of the K studies, $i = 1,2,....,K$, in the review and can represented as

$$g_i^C = \frac{(\bar{Y}_i^C - \bar{X}_i^C)}{S_i^C} \quad \text{and} \quad g_i^U = \frac{(\bar{Y}_i^U - \bar{X}_i^U)}{S_i^U},$$

where $\bar{X}_i^C$ and $\bar{X}_i^U$ represent the pretest SAT means for the coached and uncoached groups. $\bar{Y}_i^C$ and $\bar{Y}_i^U$ represent the posttest SAT means for coached and uncoached groups respectively. S_i^C and S_i^U represent their respective pretest standard deviations. For each of the two SAT subtests, separate standardized mean change measures were computed for coached and uncoached groups.

In this review, the unbiased estimates of standarized mean changes are calculated for coached and uncoached groups for both SAT-V and SAT-M subtests. The unbiased estimates of the coached and uncoached standardized mean changes are

$$d_i^C = \frac{4(n_i^C - 2)}{4n_i^C - 5} \left(\frac{\bar{Y}_i^C - \bar{X}_i^C}{S_i^C} \right),$$

and

$$d_i^U = \frac{4(n_i^U - 2)}{4n_i^U - 5} \left(\frac{\bar{Y}_i^U - \bar{X}_i^U}{S_i^U} \right),$$

where n_i^C and n_i^U are the coached and uncoached groups sample sizes.

The estimated variances of d_i^C and d_i^U are

$$Var(d_i^C) = \frac{4(1 - r_{XY}^C) + (d_i^C)^2}{2n_i^C},$$

and

$$Var(d_i^U) = \frac{4(1 - r_{XY}^U) + (d_i^U)^2}{2n_i^U}.$$

The coaching effect-size, $\hat{\Delta}_i$, is the difference between the coached and uncoached unbiased standardized mean change measures for each of the SAT subtests within each of the K studies and is denoted as

$$\hat{\Delta}_i = d_i^C - d_i^U .$$

Thus, studies that examine the effects of coaching on both SAT-M and SAT-V will have two effect sizes ($\hat{\Delta}_i^V$ for the SAT-V standardized mean-change difference and $\hat{\Delta}_i^M$ for the SAT-M subtest).

The estimated variance of $\hat{\Delta}_i$ is calculated as follow

$$Var(\hat{\Delta}_i) = \frac{4(1 - r_{XY}^C) + (d_i^C)^2}{2n_i^C} + \frac{4(1 - r_{XY}^U) + (d_i^U)^2}{2n_i^U},$$

where r_{XY}^C and r_{XY}^U are the estimators of the pretest-posttest correlations for the coached and uncoached groups respectively. In this review, the value of 0.88 was used to represent the pretest-posttest correlation for coached and uncoached samples for both SAT-V and SAT-M (DerSimonian and Laird, 1983).

The estimated covariance between $\hat{\Delta}_i^V$ and $\hat{\Delta}_i^U$ is calculated as

$$Cov(\hat{\Delta}_i^V, \hat{\Delta}_i^M) = r_{VM} [\sqrt{V(d_i^{CV}) V(d_i^{CM})} + \sqrt{V(d_i^{UV}) V(d_i^{UM})}],$$

where r_{VM} is the correlation between the SAT-Math and the SAT-Verbal subtests. Here we used the value of 0.66 to represent this correlation (Kalaian & Becker, 1986).

Thus, the estimated variance-covariance matrix for each study can be represented as

$$\underline{\hat{\sigma}}_i = \begin{bmatrix} Var(\hat{\Delta}_i^V) & Cov(\hat{\Delta}_i^V, \hat{\Delta}_i^M) \\ Cov(\hat{\Delta}_i^V, \hat{\Delta}_i^M) & Var(\hat{\Delta}_i^M) \end{bmatrix}.$$

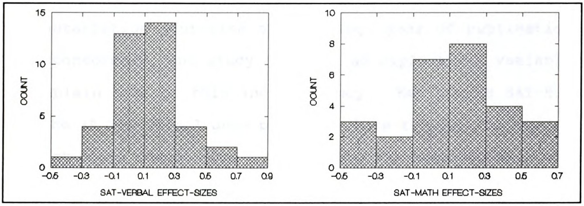
Finally, these multiple coaching effect sizes are analyzed and modeled by utilizing a multivariate mixed-effects model for meta-analysis outlined in Chapter 4. In this conceptualization, the multiple effect sizes are viewed as varying randomly across the different coaching studies and the variation among the multiple coaching effect size is modeled simultaneously as a function of study characteristics plus random error. Thus, this procedure allows one to have two effect sizes (SAT-V and SAT-M) from some of the studies as well as single effect sizes (either SAT-V or SAT-M) from the rest of the studies in the review.

5. RESULTS

The values of the SAT-Verbal effect sizes for the 39 samples ranged from -0.35 to 0.72 in standard deviation units with an overall weighted average coaching effect of 0.12 and standard deviation 0.22, while the 28 SAT-Math effect sizes ranged from -0.49 to 0.60 with an overall weighted average of 0.11 and standard deviation 0.28 (Figure 1). Thus, the average effect of coaching on SAT-Verbal and SAT-Math gains appear to be quite similar. Note that the SAT-Math average effects are smaller than in previous reviews but the SAT-Verbal effects are about the same. Although most of the coaching effect sizes are positive, the magnitudes of the coaching effects appear quite variable for both subtests.

FIGURE 1

Frequency Distribution of SAT Effect Sizes



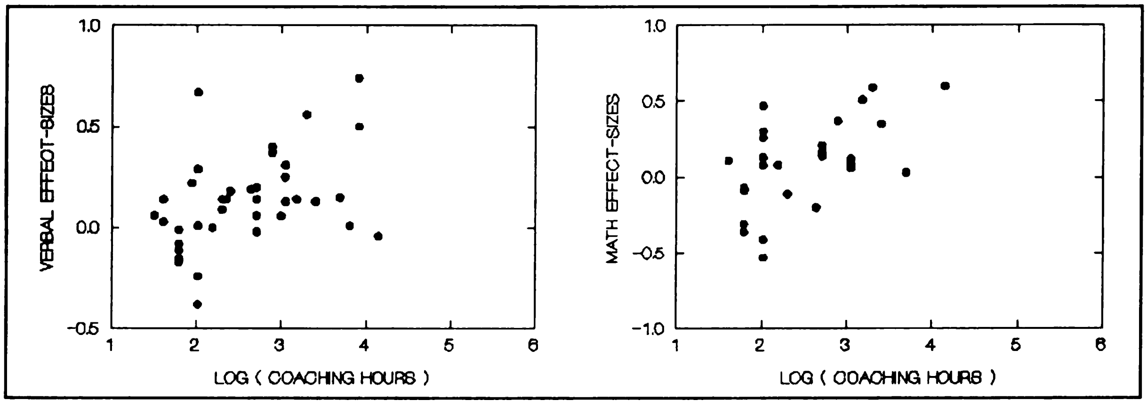
The first question I tried to answer with the application of the Multivariate Hierarchical Linear Model to SAT coaching studies was the degree of consistency of the effect sizes for both subtests and the multivariate empirical Bayes estimates of the average of these effect sizes. The results of fitting an unconditional hierarchical linear regression model (Table 5) show that the average SAT-Verbal effect size is 0.12 (p < 0.000) and it is 0.13 (p < 0.004) for SAT-Math. Contrasting the SAT-Verbal and SAT-Math regression coefficients show that

there is no significant differences between the two coefficients. Furthermore, the findings show that SAT-Math samples are more variable than SAT-Verbal samples ($\tau_V^2 = 0.006, p = 0.027$ vs $\tau_M^2 = 0.03$, $p = 0.000$).

As a result of the inconsistency in the coaching effects for both subtests across the studies, I consider study characteristics (duration of coaching, year of publication, ETS sponsorship, and study quality) as explanatory variables to explain some of this inconsistency. For the 28 SAT-Math and the 39 SAT-Verbal data points in this review, the student contact hours ranged from 4 to 93 hours for both subtests with an average coahing hours of 15 for SAT-Math and 17 for SAT-Verbal and most of the data points are clustered at the low end of the number of hours dimension (see Table 6). For this reason and the fact that there are diminishing returns in both SAT subtests' scores (Messick and Jungeblut, 1981), we used in the analysis the logarithmic tranformation of the hours of coaching dimension. Logarithmically transformed contact hours provides more accurate representation of the functional relationships between coaching effect and the number of coaching hours. The results show that SAT coaching effect sizes is moderlately related (0.5 for SAT-Verbal and 0.4 for SAT-Math) to duration of contact hours (see Figure 2).

FIGURE 2

Relationships between SAT Effect Sizes and Log (Contact Time)



The results of fitting conditional hierarchical linear model (Table 7) show that the logarithmically transformed duration of coaching has a significant positive effect on SAT-Math coaching effect sizes even after controlling for other variables in the model (β = 0.15, p = 0.04). As we can see in Table 7, no other variables studied in this review had a significant effect on SAT scores. Also, the results show that after accounting for some of the study characteristics, still considerable and significant variability left in the coaching effect sizes (τ^2_V = 0.008, p = 0.03 and τ^2_M = 0.03, p = 0.000).

Furthermore, the results show that the estimated covariance between SAT-Verbal and SAT-Math effect sizes is about -0.01.

6. FIXED-AND-MIXED-EFFECTS MODELS COMPARED

Although the fixed-effects approach is statistically developed (Raudenbush, Becker, and Kalaian, 1988; Gleser and Olkin, 1994), the actual analytical procedure is complex and needs a special computer skills from meta-analysts in order to perform a meta-analytic review. Thus, in this section, the multivariate fixed-effects model is carried out by applying multiple regression analysis, using the available standard statistical packages (SPSS-PC, SAS, SYSTAT,...etc.), on the transformed GLS within-study model which is developed in Chapter 4. Additionally, the parameter estimates of this application (multivariate fixed-effects model) to SAT coaching data set is compared to the parameter estimates obtained from applying the multivariate mixed-effects model which is developed in Chapter 4.

From the findings of the application of the multivariate mixed-effects model to SAT coaching data in the previous section, I learned that duration of coaching was the only significant explanatory varaiable. Thus, for comparison

purposes, the number of coaching hours is considered in this section as predictor variable in the model.

The results of fitting the conditional multivariate mixed-effects model show that the logarithmically transformed duration of coaching has a significant positive effect on SAT-Math coaching effect sizes (Table 8). On the other hand, the results of fitting the conditional multivariate-fixed model (Table 8) show that the logarithmically transformed coaching hours is not statistically significant. Also, from these results, we can see that the multivariate fixed-effects model yielded standard errors for the beta coefficients smaller than the mixed effects model.

7. DISCUSSION

The results of the multivariate hierarchical linear model for coaching effect sizes showed that both SAT coaching programs, on average, had positive effects of about 0.11 of a standard deviation or about six points for both SAT-Verbal and SAT-Math scores. Also, the results indicated that the average SAT-Verbal effect sizes is not significantly different from the average SAT-Math effect sizes. However, although we found great variability for the effects of coaching for both subtests, the coaching effects for SAT-Math were more variable than the SAT-Verbal coaching effects. When we modeled the variability of the effect sizes as a function of study features, student contact hours was the only significant predictor (especially for SAT-Math effect sizes) even after we controlled for the other predictors in the model. This result agrees with the previous findings of Messick and Jungeblut (1981) and Kalaian and Becker (1986) who found that duration of coaching had a strong effect on SAT scores. I also discovered that the design of the study, the publication year, and whether or not the coaching program is sponsored by Educational Testing Service did not have significant effects in explaining the variability in coaching studies.

In comparing the results of analyzing the SAT coaching effect sizes using the multivariate mixed-effects model and the multivariate fixed-effects model, the logarithmically transformed coaching hours yielded significant positive effect on SAT-Math effect sizes using the multivariate mixed-effects model. These results prove the existance of parameter variability in the coaching studies that should be accounted for by using the mixed-effects models.

Table 3

Effect Sizes of SAT Coaching Studies

Study	Year	n^c	n^u	$\hat{\Delta}^V$	$\hat{\Delta}^M$	Hours	ETS	Study Type	Home Work
				Randomized Studies					
Alderman & Powers (A)	1980	28	22	0.22	.	7	1	1	1
Alderman & Powers (B)	1980	39	40	0.09	.	10	1	1	1
Alderman & Powers (C)	1980	22	17	0.14	.	10.5	1	1	1
Alderman & Powers (D)	1980	48	43	0.14	.	10	1	1	1
Alderman & Powers (E)	1980	25	74	-0.01	.	6	1	1	1
Alderman & Powers (F)	1980	37	35	0.14	.	5	1	1	1
Alderman & Powers (G)	1980	24	70	0.18	.	11	1	1	1
Alderman & Powers (H)	1980	16	19	0.01	.	45	1	1	1
Evans & Pike (A)	1973	145	129	0.13	0.12	21	1	1	1
Evans & Pike (B)	1973	72	129	0.25	0.08	21	1	1	1
Evans & Pike (C)	1973	71	129	0.31	0.09	21	1	1	1
Laschewer	1986	13	14	0.00	0.08	8.9	0	1	0
Roberts & Oppenheim (A)	1966	43	37	0.01	.	7.5	1	1	0
Roberts & Oppenheim (B)	1966	19	13	0.67	.	7.5	1	1	0
Roberts & Oppenheim (D)	1966	16	11	-0.66	.	7.5	1	1	0
Roberts & Oppenheim (E)	1966	20	12	-0.21	.	7.5	1	1	0
Roberts & Oppenheim (F)	1966	39	28	0.31	.	7.5	1	1	0
Roberts & Oppenheim (G)	1966	38	25	.	0.26	7.5	1	1	0
Roberts & Oppenheim (H)	1966	18	13	.	-0.41	7.5	1	1	0
Roberts & Oppenheim (I)	1966	19	13	.	0.08	7.5	1	1	0
Roberts & Oppenheim (J)	1966	37	22	.	0.30	7.5	1	1	0
Roberts & Oppenheim (K)	1966	19	11	.	-0.53	7.5	1	1	0
Roberts & Oppenheim (L)	1966	17	13	.	0.12	7.5	1	1	0
Roberts & Oppenheim (M)	1966	20	12	.	0.26	7.5	1	1	0
Roberts & Oppenheim (N)	1966	20	13	.	0.47	7.5	1	1	0
Zuman (B)	1988	16	17	0.14	0.51	24	0	1	1

Table 3 (cont)
Effect Sizes of SAT Coaching Studies

Study	Year	n^C	n^U	$\hat{\Delta}^V$	$\hat{\Delta}^M$	Hours	ETS	Study Type	Home Work
				Matched Studies					
Burke (A)	1986	25	25	0.50	.	50	0	2	1
Burke (B)	1986	25	25	0.74	.	50	0	2	1
Coffin (A)	1987	8	8	-0.20	0.37	18	0	2	0
Davis	1985	22	21	0.14	-0.14	15	0	2	0
Frankel	1960	45	45	0.13	0.35	30	0	2	0
Kintisch	1979	38	38	0.14	.	20	0	2	1
Whitla	1962	52*	52*	0.09	-0.11	10	1	2	1
				Nonequivalent Comparison Studies					
Curran (A)	1988	21	17	.	.	6	0	3	0
Curran (B)	1988	24	17	.	.	6	0	3	0
Curran (C)	1988	20	17	.	.	6	0	3	0
Curran (D)	1988	20	17	.	.	6	0	3	0
Dear	1958	60	526	-0.02	0.21	15	1	3	1
Dyer	1953	225	193	0.06	0.27	15	1	3	1
French (B)	1955	110	158	0.06	.	4.5	1	3	1
French (C)	1955	161	158	.	0.20	15	1	3	1
FTC (A)	1978	192	684	0.34	0.31	40	0	3	0
Keefauver	1976	16	25	0.19	-.20	14	0	3	0
Lass	1961	38	82	0.03	0.11	.	1	3	1
Reynolds & Oberman	1987	93	47	-0.04	0.59	63	0	3	1
Teague	1992	10	15	0.40	.	18	0	3	0
Zuman (A)	1988	21	34	0.56	0.59	27	0	3	1

* The sample sizes for SAT-V were $n^C = 52$ and $n^U = 52$.

Table 4

Characteristics and Features of SAT Coaching Studies

Characteristic	Coded Values	
Randomized Study	(1) yes	(0) no
Student Voluntariness	(1) yes	(0) no
Presence of Verbal Coaching	(1) yes	(0) no
Presence of Math Coaching	(1) yes	(0) no
Assignment of Homework	(1) yes	(0) no
ETS Sponsored Research	(1) yes	(0) no
Publication Year	last two digits of the year	
Coaching Duration	log (hours)	

Table 5

Frequency Distribution of Student Contact Hours

Categories (in hours)	SAT-V Samples	SAT-M Samples
4.5 - 10	18	15
10.5 - 20	10	6
20.5 - 30	6	6
30.5 - 40	1	1
40.5 - 50	3	0
> 50.5	1	1
Mean	17.2	15.4
S. D.	14.4	12.8
Total	39	28

Table 6

Fitting Unconditional Model Results

Fixed and Random Effects	Coefficient	Standard Error	t-ratio	P-value
For SAT-V				
Intercept	0.118	0.021	5.51	0.00
τ^2 - estimate	0.006			
For SAT-M				
Intercept	0.125	0.039	3.18	0.004
τ^2 - estimate	0.03			

Table 7

Fitting Conditional Model Results

Fixed and Random Effects	Coefficient	Standard Error	t-ratio	P-value
For SAT-V				
Intercept	0.099	0.049	2.06	0.06
Year	0.002	0.004	0.48	0.39
Log (hours)	0.075	0.002	1.94	0.13
ETS	0.079	0.118	0.68	0.36
Randomized	0.003	0.089	0.03	0.38
τ^2_V - estimate	0.008			
For SAT-M				
Intercept	0.057	0.32	0.77	0.29
Year	-0.000	0.39	-0.19	0.39
Log (hours)	0.15	0.04	2.47	0.02
ETS	-0.016	0.39	-0.12	0.39
Randomized	0.07	0.34	0.63	0.32
τ^2_M - estimate	0.03			

Table 8

COMPARISON BETWEEN FIXED-AND-MIXED-EFFECTS

MODEL ESTIMATES

VARIABLES	Mixed		Fixed	
	$\hat{\beta}$	$S(\hat{\beta})$	$\hat{\beta}$	$S(\hat{\beta})$
For SAT-V				
- Intercept	-0.04	0.09	-0.05	0.05
- LOG (Coaching Hours)	0.06	0.03	0.05	0.02
For SAT-M				
- Intecept	-0.27	0.17	-0.23	0.15
- LOG (Coaching Hours)	0.15 *	0.06	0.07	0.05

* Indicates Significance at $p < .05$

CHAPTER VIII

DISCUSSION AND IMPLICATIONS

In preceding chapters, the multivariate mixed-effects model was first developed. Second, the empirical Bayes estimates of the parameters for the model were derived. Finally, the applicability of the proposed model to artificial and real data sets was illustrated. Additionally, the parameter estimates from applying the multivariate mixed-effects and the multivariate fixed-effects models were compared. Although the concluding statements about these analyses were provided in the previous two chapters, some important conclusions will be restated in this chapter.

I learned from the application of the V-Known routine and the Hierarchical Linear Model (HLM) program to the artificial data set that the HLM program can be used instead of the V-Known routine for research-synthesis purposed to obtain empirical Bayes parameter estimates using the multivariate mixed-effects model. Since the proposed model can be used to model effect-size data with missing values, the HLM program can be used to analyze multiple correlated effect sizes for each study in the meta-analysis with missing effect sizes from some of the studies.

The results of applying the multivariate mixed-effects model to Scholastic Aptitude Test (SAT) coaching effects studies showed that the HLM program can be applied appropriately to multivariate meta-analysis with missing effect-size data points. This evidence of the applicability of the HLM program for analyzing multivariate effect-size data with or without missing effect sizes can help us to carry more design-oriented meta-analyses. For instance, we can use HLM to take into account and incorporate within-study characteristics in the multivariate mixed-effects model.

Another significant contribution of the proposed model (multivariate mixed effects model) is its practical use to perform multivariate fixed-effects model statistical analysis. I illustrated the use of the proposed multivariate mixed-effects model to obtain multivariate fixed-effects parameter estimates by using standard statistical computer packages as well as multivariate mixed-effects parameter estimated by using the HLM computer software.

Given the importance and the seriousness of the "missing effect-sizes problem" in meta-analysis and research synthesis, the effects of missing effect sizes in multivariate data sets should be further explored and examined more closely using simulation studies. Also, the behavior of the empirical Bayes estimates when specific percentages (e.g. 5%, 10%, 15% and

25%) of the effect sizes are missing should be further studied.

In this study, the application of the HLM program and the proposed multivariate mixed-effects model was illustrated using bivariate artificial and real data sets. As substantive future research, the application of the illustrated methodology should be applied to meta-analysis studies with more than two outcomes, with or without missing effect sizes. Also, these new applications should consider taking into account the within-study characteristics and incorporating them in the multivariate mixed-effects model. Furthermore, the robustness of violating the assumptions of the proposed model should be studied.

APPENDICES

APPENDIX A

V - KNOWN COMPUTER OUTPUT

```
*********************************************************
*                                      *
*     H   H L     M   M  22               *
*     H   H L     MM MM  2  2               *
*     HHHHH  L      M M M    2     Version 3.01    *
*     H   H L     M   M  2                *
*     H   H LLLLL  M   M  2222               *
*                                      *
*********************************************************
```

SPECIFICATIONS FOR THIS HLM RUN   Sat May 28 11:44:50
1994

Problem Title: Multivariate HLM for Generated Data (No Predictors)

The data source for this run = sim1.ssm
Output file name   = output
The maximum number of level-2 units = 50
The maximum number of iterations = 3000
Weighting Specification

	Weighting?	Weight Variable Name	Normalized?
Level 1	no		no
Level 2	no		no

The outcome variable is   EF

The model specified for the fixed effects was:

Level-1 Coefficients	Level-2 Predictors
V slope, B1	INTRCPT2, G10
M slope, B2	INTRCPT2, G20

The model specified for the covariance components was:

Sigma squared (constant across level-2 units)

Tau dimensions
V slope
M slope

Summary of the model specified (in equation format)

Level-1 Model

Y = B1*(V) B2*(M) + R

Level-2 Model

B1 = G10 + U1
B2 = G20 + U2

Level-1 OLS regressions

Level-2 Unit	V slope	M slope

The average OLS level-1 coefficient for V = 0.02581
The average OLS level-1 coefficient for M = -0.03986

STARTING VALUES
----
sigma(0)_squared = 1.00000

Tau(0)

V	0.03319	0.00707
M	0.00707	0.02643

The outcome variable is EF

Estimation of fixed effects
(Based on starting values of covariance components)

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
For V slope, B1				
INTRCPT2, G10	0.037941	0.035791	1.060	0.225
For M slope, B2				
INTRCPT2, G20	-0.021686	0.033678	-0.644	0.321

The value of the likelihood function at iteration 1 = -1.989962E+002

The value of the likelihood function at iteration 2 = -1.982290E+002

The value of the likelihood function at iteration 3 = -1.978401E+002

The value of the likelihood function at iteration 4 = -1.976544E+002

The value of the likelihood function at iteration 5 = -1.975194E+002

.
.
.

The value of the likelihood function at iteration 7 = -1.975023E+002

The value of the likelihood function at iteration 8 = -1.975010E+002

The value of the likelihood function at iteration 9 = -1.975006E+002

The value of the likelihood function at iteration 10 = -1.975004E+002

Iterations stopped due to small change in likelihood function
******* ITERATION 11 *******

Sigma_squared = 1.00000

Tau

V	0.05411	0.02010
M	0.02010	0.03614

Tau (as correlations)

V	1.000	0.454
M	0.454	1.000

Random level-1 coefficient	Reliability estimate
V, B0	0.622
M, B1	0.535

The value of the likelihood function at iteration 11 = -1.975004E+002

The outcome variable is EF

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
For V slope, B1				
INTRCPT2, G10	0.035176	0.041652	0.845	0.276
For M slope, B2				
INTRCPT2, G20	-0.023977	0.036768	-0.652	0.320

Final estimation of variance components:

Random Effect	Standard Deviation	Variance Component	df	Chi-square	P-value
V slope, U0	0.23262	0.05411	49	153.90418	0.000
M slope, U1	0.19011	0.03614	49	116.82539	0.000
level-1, R	1.00000	1.00000			

Statistics for current covariance components model

Deviance = 395.00079
Number of estimated parameters = 4

APPENDIX B

HLM COMPUTER OUTPUT

```
*********************************************************
*                                       *
*    H  H L    M  M  22               *
*    H  H L    MM MM  2  2              *
*    HHHHH  L     M M M   2    Version 3.01   *
*    H  H L    M  M  2              *
*    H  H LLLLL  M   M  2222              *
*                                       *
*********************************************************
```

SPECIFICATIONS FOR THIS HLM RUN   Tue May 31 10:59:04
1994

Problem Title: Multivariate V-Known for Generated Data (No Predictors)

The data source for this run = c:\dis\d1.ssm
Output file name   = output1
The maximum number of level-2 units = 50
The maximum number of iterations = 3000
Note: this is a v-known analysis

The model specified for the fixed effects was:

Level-1 Effects	Level-2 Predictors
V, B1	INTRCPT2, G10
M, B2	INTRCPT2, G20

The model specified for the covariance components was:

Variance(s and covariances) at level-1 externally specified

Tau dimensions
V
M

Summary of the model specified (in equation format)

Level-1 Model

Y1 = B1 + E1
Y2 = B2 + E2

Level-2 Model

B1 = G10 + U1
B2 = G20 + U2

STARTING VALUES

Tau(0)

V	0.03263	0.00659
M	0.00659	0.02602

Estimation of fixed effects
(Based on starting values of covariance components)

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
INTRCPT2, G10	0.038054	0.035615	1.068	0.223
INTRCPT2, G20	-0.021578	0.033540	-0.643	0.322

The value of the likelihood function at iteration 1 = -2.144441E+002

The value of the likelihood function at iteration 2 = -2.128138E+002

The value of the likelihood function at iteration 3 = -2.119697E+002

The value of the likelihood function at iteration 4 = -2.115597E+002

The value of the likelihood function at iteration 5 = -2.112609E+002

.
.
.

The value of the likelihood function at iteration 7 = -2.112178E+002

The value of the likelihood function at iteration 8 = -2.112146E+002

The value of the likelihood function at iteration 9 = -2.112135E+002

The value of the likelihood function at iteration 10 = -2.112129E+002

Iterations stopped due to small change in likelihood function
******* ITERATION 11 *******

Tau

V	0.05411	0.02010
M	0.02010	0.03615

Tau (as correlations)

V	1.000	0.454
M	0.454	1.000

Random level-1 coefficient	Reliability estimate
V, B1	0.622
M, B2	0.535

The value of the likelihood function at iteration 11 = -2.112129E+002

Final estimation of fixed effects:

Fixed Effect	Coefficient	Standard Error	T-ratio	P-value
INTRCPT2, G10	0.035177	0.041652	0.845	0.276
INTRCPT2, G20	-0.023977	0.036769	-0.652	0.320

Final estimation of variance components:

Random Effect		Standard Deviation	Variance Component	df	Chi-square	P-value
V,	U1	0.23261	0.05411	49	153.90286	0.000
V,	U2	0.19012	0.03615	49	116.82497	0.000

Statistics for current covariance components model

Deviance = 422.42582
Number of estimated parameters = 4

REFERENCES

Alderman, D. L., & Powers, D. E. (1980). The effects of special preparation on SAT-verbal scores. *American Educational Research Journal, 17*, 239-253.

Becker, B. J. (1986). Influence again: An examination of reviews and studies of gender differences in social influence. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis* (pp. 178-209). Baltimore, MD: Johns Hopkins University Press.

Becker, B. J. (1988). Synthesizing standardized mean-change measures. *British Journal of Mathematical and Statistical Psychology, 41*, 257-278.

Becker, B. J. (1990). Coaching for the scholastic aptitude test: Further synthesis and appraisal. *Review of Educational Research, 60* (3), 373-417.

Berlin, J. A., & Colditz, G. A. (1990). A meta-analysis of physical activity in the prevention of coronary heart disease. *American Journal of Epidemiology, 132*, 612-628.

Burke, K. B. (1986). *A model reading course and its effects on the verbal scores of eleventh and twelfth grade students on the Nelson Denny Test, the Preliminary Scholastic Aptitude Test, and the Scholastic Aptitude Test*. Unpublished doctoral dissertation, Georgia State University. (University Microfilms No. 86-26152).

Bryk, A. S., Raudenbush, S. W., Seltzer, M. & Congdon, R. T. (1986). *An Introduction to HLM: Computer Program and Users' Guide*. (Available from A. S. Bryk, Department of Education, University of Chicago, 5835 South Kimbark, Chicago, IL 60637).

Coffin, G. C. (1987). *Computer as a tool in SAT preparation*. Paper presented at the annual meeting of the Florida Instructional Computing Conference, Orlando, FL. (ERIC Document Reproduction Service No. ED 286 932).

Coffman, W. E., & Parry, M. E. (1967). Effects of an accelerated reading course on SAT-V scores. *Personnel and Guidance Journal, 46*, 292-296.

Comras, J. (1984). The SAT - What does it measure and does it still work? *NASSP Bulletin,* 50-59.

Curran, R. G. (1988). *The effectiveness of computerized coaching for the Preliminary Scholastic Aptitude Test (PSAT/NMSQT) and the Scholastic Aptitude Test (SAT).* Unpublished doctoral dissertation, Boston University. (University Microfilms No. 88-14377).

Davis, W. D. (1985). *An empirical assessment of selected computer software purported to raise SAT scores significantly when utilized with short-term computer-assisted instruction on the microcomputer.* Unpublished doctoral dissertation, Florida State University. (ERIC Document Reproduction Service No. ED 283 370).

Dear, R. E. (1958). *The effect of a program of intensive coaching on SAT scores* (ETS RB 58-5). Princeton, NJ: Educational Testing Service.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B, 39*, 1-38.

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

DerSimonian, R., & Laird, N. M. (1983). Evaluating the effect of coaching on SAT scores: A meta-analysis. *Harvard Educational Review, 53* (1), 1-15.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*, 177-188.

Dyer, H. S. (1953). Does coaching help? *College Board Review, 19*, 331-335.

Dyer, H. S. (1987). The effects of coaching for scholastic aptitude. *NASSP Bulletin,* 46-50.

Evans, F. R., & Pike, L. W. (1973). The effects of instruction for three mathematics item formats. *Journal of Educational Measurement, 10*, 257-272.

Federal Trade Commission, Boston Regional Office. (1978). *Staff memorandum of the Boston Regional Office of the Federal Trade Commission: The effects of coaching on standardized admission examinations* (NTIS No. PB-296 210). Boston: Author.

Federal Trade Commission, Bureau of Consumer Protection. (1979). *Effects of coaching on standardized admission examinations: Revised statistical analyses of data gathered by Boston Regional Office of the Federal Trade Commission* (NTIS No. PB-296 196). Washington, DC: Author.

Finn, J. D. (1974). *A General Model for Multivariate Analysis*. Holt, Rinehart, and Winston, Inc.

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron, 1*, 1-32.

Frankel, E. (1960). Effects of growth, practice, and coaching on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 38*, 713-719.

French, J. W. (1955). *The coachability of the SAT in public schools* (ETS RB 55-26). Princeton, NJ: Educational Testing Service.

French, J. W., & Dear, R. E. (1959). Effects of coaching on an aptitude test. *Educational and Psychological Measurement, 19,* 319-330.

Giaconia, R. M., & Hedges, L. V. (1982). Identifying features of effective open education. *Review of Educational Research, 52*, 579-602.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*, 3-8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Gleser, L. J., & Olkin, I. (1994). Stochastically dependent effect sizes. In Cooper, H. & Hedges, L. V. (Eds.), *The Handbook of Research Synthesis.* New York: Russell Sage Foundation.

Greenland, S. (1993). A Meta-analysis of coffee, myocardial infarction, and coronary death. *Epidemiology, 4* (4), 366-374.

Greenland, S. (1987). Quantitative methods in the review of epidemiologic literature. *Epidemiology Review, 107*, 224-233.

Hedges, L. V. (1981). Distribution theory for Glass's estimator for effect size and related estimators. *Journal of Educational Statistics, 6* (2), 107-128.

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

Hedges, L. V. (1982b). Fitting continuous models to effect size data. *Journal of Educational Statistics, 7* (4), 245-270.

Hedges, L. V. (1983). A random effects model for effect size. *Psychological Bulletin, 93*, 388-395.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis.* New York: Academic Press.

Hugue, M.F. (1988). Experiences with meta-analysis in NDA submissions. *Proceedings of the Biopharmaceutical Section of the American Statistical Association, 2*, 28-33.

Iaffaldano, M. T., & Muchinsky, P. M. (1985). Job satisfaction and job performance: A meta-analysis. *Psychological Bulletin, 97*, 251-273.

Jenicek, M. (1989). Meta-analysis in medicine: Where we are and where we want to go. *Journal of Clinical Epidemiology, 42*, 35-44.

Johnson, S. T. (1984). *Preparing Black students for the SAT -- Does it make a difference?* (An evaluation report of the NAACP Test Preparation Project). Unpublished report to the National Association for the Advancement of Colored People, New York. (ERIC Document Reproduction Service No. ED 247 350).

Kalaian, H. A., & Becker, B. J. (1986). *The effects of coaching on Scholastic Aptitude Test (SAT) performance: A multivariate meta-analysis approach.* Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Keefauvaer, L. W. (1976). *The effects of a program of coaching on Scholastic Aptitude Test scores of high school seniors pretested as juniors.* Unpublished doctoral dissertation, University of Tennessee at Knoxville. (University Microfilms No. 77-3651).

Kintisch, L. S. (1979). Classroom techniques for improving Scholastic Aptitude Test scores. *Journal of Reading, 22*, 416-419.

Kulik, J. A., Bangert-Drowns, R. L., & Kulik, C. L. C. (1984). Effectiveness of coaching for aptitude tests. *Psychological Bulletin, 95*, 179-188.

L'Abbee, K. A., Detsky, A. S., & Rouke, K. (1987). Meta-analysis in clinical research. *Annals of Internal Medicine, 107*, 224-233.

Laschewer, A. D. (1986). *The effect of computer assisted instruction as a coaching technique for the Scholastic Aptitude Test preparation of high school juniors.* Unpublished doctoral dissertation, Hofstra University. (University Microfilms No. 86-06936).

Lass, A. H. (1961). Unpublished report. (Cited in Pike, L.W. (1978). *Short-term instruction, testwiseness, and the Scholastic Aptitude Test: A literature review with recommendations.* New York: College Entrance Examination Board.)

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

Little, R. J. A. (1988a). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Little, R. J. A. (1988b). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistician, 37*, 23-38.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis With Missing Data.* New York: Wiley.

Little, R. J. A., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods and Research, 18*, 292-326.

Longnecker, M. P., Berlin, J. K., Orza, M. J., & Chlamers T. C. (1988). A meta-analysis of alcohol consumption in relation to risk of breast cancer. *Journal of American Medical Association, 260*, 652-656.

Marron, J. E. (1965). *Preparatory school test preparation: Special test preparation, its effect on College Board scores and the relationship of affected scores to subsequent college performance.* West Point: Research Division, Office of the Director of Admissions and Registrar, United States Military Academy.

Messick, S., & Jungeblut, A. (1981). Time and method in coaching for the SAT. *Psychological Bulletin, 89,* 191-216.

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

Pallone, N. J. (1961). Effects of short-term and long-term developmental reading courses upon SAT verbal scores. *Personnel and Guidance Journal, 39*, 654-657.

Pigott, T. D. (1992). *The application of maximum likelihood methods to missing data in meta-analysis.* Unpublished Doctoral Dissertation, University of Chicago.

Pigott, T. D. (1994a). Methods for handling missing data in research synthesis. In Cooper, H. and Hedges, L. V. (Eds.), *The Handbook of Research Synthesis.* Russell Sage Foundation.

Pigott, T. D. (1994b). *The application of maximum likelihood methods for missing data in meta-analysis.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Ottenbacher, K. P., & Petersen, P. (1983). Quantitative reviewing of medical literature: An approach to synthesizing research results in clinical procedures. *Clinical Pediatric, 28*, 423-427.

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

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

Raudenbush, S. W., Becker, B. J., & Kalaian, H. (1988). Modeling multivariate effect sizes. *Psychological Bulletin, 103*, 111-120.

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

Reynolds, A. J., & Oberman, G. O. (1987). *An analysis of a PSAT preparation program for urban gifted students.* Paper presented at the annual meeting of the American Educational Research Association, Washington, DC.

Roberts, S. O., & Oppenheim, D. B. (1966). *The effect of special instruction upon test performance of high school students in Tennessee* (CB RDR 66-7, No. 1, and ETS RB 66-36). Princeton, NJ: Educational Testing Service.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences, 3*, 377-386.

Rosenthal, R., & Rubin, D. B. (1986). Meta-analytic procedures for combining studies with multiple effect sizes. *Psychological Bulletin, 99*, 400-406.

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

Sacks, H. S., et al (1987). Meta-analyses of randomized controlled trials. *New England Journal of Medicine, 316*, 450-455.

Shinton, R., & Beevers, G. (1989). Meta-analysis of relation between cigarette smoking and stroke. *British Medical Journal, 298*, 789-794.

Slack, W. V., & Porter, D. (1980). The Scholastic Aptitude Test: A critical appraisal. *Harvard Educational Review, 50,* 154-175.

Teague, A. B. (1992). *A program to help college bound students improve their verbal Scholastic Aptitude Test scores*. Unpublished master's thesis, Nova University. (ERIC Document Reproduction Service No. CS 213 534).

Thacker, S. B. (1988). Meta-analysis: A quantitative approach to research integration. *Journal of American Medical Association, 259*, 1685-1689.

White, K. R. (1982). The relation between socioeconomic status and academic achievement. *Psychological Bulletin, 91* (3), 461-481.

Whitla, D. K. (1962). Effect of tutoring on Scholastic Aptitude Test scores. *Personnel and Guidance Journal, 41,* 32-37.

Yusuf, S., Peto, R., Lewis, J., Collins, R., & Sleight, P. (1987). Beta-blockade during and after myocardial infarction: an overview of the randomized trials. *New England Journal of Medicine, 316,* 450-455.

Zuman, J. P. (1988). *The effectiveness of special preparation for the SAT: An evaluation of a commercial coaching school.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans. (ERIC Document Reproduction Service No. ED 294 900).

MICHIGAN STATE UNIV. LIBRARIES
31293010204976