This is to certify that the

dissertation entitled

An Investigation of Methods
for Mixed-Model Meta-Analysis
in the Presence of Missing Data

presented by

Kyle R. Fahrbach

has been accepted towards fulfillment
of the requirements for

_____ Ph.d. ____ degree in __CEPSE_____

Betsy Jane Becker
Major professor

Date _____5-9-01_____

135 210 THS
Main Wgrt

PLACE IN RETURN BOX to remove this checkout from your record.
TO AVOID FINES return on or before date due.
MAY BE RECALLED with earlier due date if requested.

DATE DUE	DATE DUE	DATE DUE
DEC 2 3 2004 0 7 0 6 0 4		
AUG 2 3 2005		

6/01 c:/CIRC/DateDue.p65-p.15

An Investigation of Methods
for Mixed-Model Meta-Analysis
in the Presence of Missing Data

By

Kyle R. Fahrbach

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

2001

ABSTRACT

An Investigation of Methods
for Mixed-Model Meta-Analyses in the
Presence of Missing Data

By

Kyle R. Fahrbach

Meta-analysts often find that the data sets they are planning to synthesize have missing data on potentially important study characteristics. When the data are complete, these study characteristics may be controlled for and their effects estimated through techniques similar to weighted-least squares multiple regression analysis (Hedges & Olkin, 1985; Raudenbush, 1994), where the study characteristics are treated as predictors and the effect-size magnitude is the outcome. When data are missing, however, new analysis techniques must be employed if the meta-analyst does not want to resort to either dropping potentially important study characteristics from the analysis, or dropping studies that are missing data on those characteristics.

In the present study I investigate the estimation of parameters in a mixed-model meta-analysis under the conditions that there is missing data on the predictors. To date only Pigott (1992) has investigated estimation in meta-analytic models where data are missing, and she did not model the presence of random effects. The estimation procedures compared here include complete-case analysis (the default in meta-analysis today), available-case analysis, and maximum-likelihood estimation through the Expectation-Maximization (EM) algorithm. Each procedure is compared with regard to the bias and efficiency of its estimators for three different types of missing data: missing completely at

random (MCAR), missing at random (MAR), and not missing at random (NMAR).

Bootstrapped standard errors for the MLE method (using the method outlined in Su, 1988) are generated and examined for accuracy, and a real meta-analytic dataset of juvenile delinquency reduction studies is examined using the methods mentioned above.

The findings show that on average, the EM maximum-likelihood estimation procedure produced substantively important gains in efficiency over both CC and AC estimation, though there were a few subsets of simulation parameters for which the improvement was either small or nonexistent.

The bootstrapped standard errors for the slopes were, on average, very accurate and had acceptable non-coverage rates. However, the bootstrapped standard errors for the estimation of τ proved to be too conservative. The estimation of multiple models for the juvenile delinquency dataset showed that the program that employs the EM estimation procedure can handle models with many variables and switch between models easily. With this program in hand, the added inconvenience of maximum-likelihood estimation in meta-analyst becomes minimal.

DEDICATION

To my father, who can now stop asking me

"Are you done yet?"

ACKNOWLEDGMENTS

To start with, I want to thank the members of my committee – Betsy Becker, Ken Frank, Mark Reckase, Richard Houang, and James Stapleton, for their guidance and support. Without it, the final version wouldn't have been as nearly as good.

Much thanks to Sam Larson. Sam, you are by far the coolest person I've ever done consulting for, and I've profited from your advice as much as your $$.

Much thanks also to Ken Frank. Ken, not only have I enjoyed and learned much by working with you, but without your referrals I'd have been living in the street for the last five years. This would have put a damper on my dissertation-writing activities.

And, of course, much, much, much, much thanks to my advisor, Betsy. I would put those "much"'s in a 32-point font, but I believe that in that form they would not pass dissertation formatting muster. Betsy, without you, I'd be wrapping up my apprenticeship proposal right about now. And it would be written in salsa, on a napkin. You've been a fantastic mentor for me through this entire program, and for that, I owe you greatly. You've also been a great friend.

Speaking of friends, I would like to acknowledge Aaron Blodgett and Jarek Hruscik. They did little to help me finish the dissertation – but they did help stop me from going insane while I was working on it.

And finally, to Anne Continelli – without me, you'd probably be far less insane. Thanks for sticking with me. Now, I shall have to find some other excuse to avoid doing housework than "But I have to work on my dissertation!"

TABLE OF CONTENTS

LIST OF TABLES

x

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Researchers conducting meta-analysis often find themselves the victim of problems for which they have little recourse. Many of these problems center around an absence of data that theoretically should be reported in the studies of the field being investigated. For instance, the problem of publication bias, otherwise known as the "file-drawer problem," concerns whether studies have gone unpublished because they failed to find significant results for whatever treatment or relationship was being studied. While methods exist to correct for this bias (see, e.g., Begg, 1994), it is a problem that is difficult to handle -- or to know the full extent of in any given field. Another common problem, and the problem that is the focus of this study, is missing data that occurs *within* studies. Missing data within studies occurs when variables that the meta-analyst believes may moderate or mediate the effect magnitude under investigation (such as a correlation or effect size) are not reported.

Missing data of this sort are uncomfortably common in primary studies. A good example is the set of studies in Lipsey's 1992 meta-analysis of juvenile-delinquency treatments. Lipsey collected 443 studies of youth delinquency interventions that had both a treatment and a control group. Fully 12% of these studies failed to report a variable as important as the average age of the juveniles during treatment, and over 25% failed to report the mean total number of hours of contact between the intervention staff and the juveniles under treatment. Missing data on important study characteristics occur

1

throughout the social sciences, and it is the meta-analyst's continuous burden to handle such deficiencies.

When confronted by missing data, researchers are generally advised to take one (or both) of two routes. They may either drop studies that have missing data on the variables they think may be important, or, to avoid lowering their sample size, they may instead drop those variables from the analysis. The first option shrinks the sample size of the meta-analysis; the second is theoretically unsound. While researchers can try to combine both routes, the effect is piecemeal at best, as the different sub-analyses work with different numbers of studies and different parts of the population of studies. A third route is sometimes taken, though it is heavily advised against in the literature: some researchers impute overall means for missing values in order that neither their overall sample size nor their variable pool is compromised. As I will show in the review of literature, this practice can lead to strongly biased estimators of meta-analytic parameters.

The problem of missing data in meta-analysis is made worse in that meta-analytic datasets are often perceived to be very diverse -- moreso than many other datasets analyzed through multiple regression. This perception has led to the "comparing apples and oranges" debate that has taken place within the meta-analysis community for the last twenty years (e.g., Green & Hall, 1984; Slavin, 1984). Given this perceived diversity and the fact that the average meta-analysis has fewer cases (i.e., studies) than the average social-science study (see, e.g., Harris & Rosenthal, 1985; Hunter & Schmidt, 1990; Lent, Auerbach, & Levin, 1971), it should be anathema to delete studies from one's analysis simply because they do not report complete data on all variables of interest.

2

As of this writing, the only researcher to examine the problem of missing data within meta-analytic studies in a statistically rigorous way has been Pigott (1992). Pigott derived maximum-likelihood (ML) estimators of regression slopes within a meta-analytic model. In this model, the outcome is some asymptotically normally distributed parameter, such as an effect size or a correlation, and the predictors are study characteristics, such as type of treatment or mean number of contact hours between juvenile and intervener. Her model was not a full model, however, in that it did not allow for the presence of a random effect for the intercept. In other words, her model assumes that after accounting for sampling error and variation in the observed predictors, all variation has been accounted for and all studies have the same population correlation or effect size. This assumption is restrictive and may be unrealistic in many fields, especially if not all important study characteristics were measured. Also, Pigott assumed that study characteristics are measured with a precision proportional to their study's sample size. Obviously, for many conceivable study characteristics (i.e., sex of author, length of treatment) this assumption is an unreasonable one; there is no reason to think that a more precise estimate of the sex of a study's author comes from a study with 20 participants than from a study with 200 or 2000. An estimation procedure that estimates a random effect, and does not make the above assumption, is the natural next step.

There is a potential problem in deriving a more complicated maximum-likelihood method, however. There is the danger that it will be too complicated or intimidating for those who generally conduct meta-analyses, i.e., subject-matter experts who may be uncomfortable or unfamiliar with complex statistical procedures. There are two possible

3

solutions to this problem: find an easier method of handling missing data, or write software that is general enough and simple enough to be used by someone whose statistical expertise is limited to moderate exposure to multiple regression and facility with one of the simpler statistical packages (e.g., SPSS for Windows).

The purpose of this study is to present, and test, a method for the estimation of meta-analytic parameters while addressing each of the above issues. A statistically rigorous maximum-likelihood estimation procedure is derived that accounts for the possible presence of a study random effect, and is written into software that can handle most meta-analytic datasets. A less statistically rigorous, but intuitively promising method for handling missing data (available-case analysis, also known as pairwise deletion) is also derived and tested against the maximum-likelihood procedure. It is expected that the maximum-likelihood procedure will perform better in terms of bias and mean-squared error (MSE) of its estimators; however, it is an open question as to whether it will perform *substantively* better. If the only difference between maximum-likelihood (ML) analysis and available-case (AC) analysis is that the MSEs of the ML estimators are only one or two percent smaller than the MSEs of the AC estimators, then practically speaking, the two techniques are performing equally. The two techniques are tested against each other and against standard complete-case analysis while varying a variety of study parameters, including study sample size, size of random effect, population correlation matrix between the predictors and outcome, number of studies, and type of missing data.

Thus, this investigation offers several important additions to the field. Meta-analytic methods have fallen behind those of related fields, such as Hierarchical Linear

Modeling (HLM) in regard to their statistical sophistication. The derivation of a maximum-likelihood estimation procedure to estimate all important meta-analytic parameters, even in the presence of missing data, brings meta-analysis back to state-of-the-art. Yet, it is not taken on faith that the new methods, which are admittedly complex, give substantively different results than less statistically rigorous methods that are easier to understand. This study quantifies both the accuracy of all the methods studied and their accuracies relative to each other, in order to determine whether the "payoff" from the use of the more complicated method is worthwhile. Finally, the development of easy to use software will allow complicated meta-analytic models to be investigated by those who are subject-matter experts and not statisticians.

Chapter 2 presents a review of the literature that defines the different types of missing data and summarizes research regarding different methods of handling different types of missing data.

Chapter 3 begins with a statement of the meta-analytic model under investigation. It then describes how complete-case analysis estimates are calculated, followed by a derivation of the available-case estimation procedure as well as the maximum-likelihood estimation method. These are followed by a summary of the simulations used to test both of these estimation procedures, as well as test the "default" estimation procedure, complete-case analysis..

Chapter 4 provides a summary of the parameters to be varied in the simulations, as well as rationales for their ranges of values.

Chapter 5 describes the criteria for investigating the different estimators of the

meta-analytic model and summarizes the results of the simulation work with regard to how biased and precise the estimators are for the different methods.

Chapter 6 presents the analysis of a real meta-analytic dataset, provided by Dr. Mark Lipsey, concerning interventions to prevent juvenile delinquency. This data set has been the source of multiple papers (e.g., Lipsey, 1999a, 1999b).

Finally, Chapter 7 summarizes the results of the simulation work and the meta-analysis of the Lipsey data set, and examines the possible substantive benefits of using the more statistically rigorous maximum-likelihood method.

CHAPTER II

REVIEW OF THE LITERATURE

Different types of missing data for the typical general linear model are introduced, followed by a description of the estimation procedures used to handle these different types of missing data. Literature (mostly simulation work) regarding the bias and efficiency of the estimators from the different procedures is summarized. While almost none of the literature that compares and contrasts the different procedures looks at a general linear model (GLM) with both a fixed error term and a random effects component, or has any sort of meta-analytic context, it helps inform us as to what procedures would be best applied to the problem.

1. Types of Missing Data

Define $Y = (Y_{ij})$ as a k x $(p+1)$ matrix of k observations measured for $(p+1)$ variables, where the first variable measured is the dependent variable and the next p variables are the independent variables.

Define the response indicator $R=(R_{ij})$, such that $R_{ij}=1$ if Y_{ij} is observed, and $R_{ij}=0$ if Y_{ij} is missing. This model treats R as a random variable; the type of missing data present will depend on the specification of Y and the distribution of R, given Y, which is

$$f(Y,R|\Phi,\psi) = f(Y|\Phi)f(R|Y,\psi). \qquad (2.2)$$

The parameter Φ represents the parameters concerning the Y_{ij} and their interrelationships. Parameter ψ represents an unknown variable or group of variables that affect the distribution of the missing-data mechanism.

We can separate the values of the Y_{ij} into two groups: Y_{obs} and Y_{mis}, where Y_{obs} denotes the observed values and Y_{mis} denotes the missing values. The observed data, then, consist of the values of the Y_{obs} as well as R. Data are considered to be missing at random (MAR) when the equality in 2.2 holds true for Y_{obs} (as opposed to Y); that is,

$$f(Y_{obs},R|\Phi,\psi) = f(Y_{obs}|\Phi)f(R|Y_{obs},\psi). \qquad \text{(MAR data)} \qquad (2.3)$$

To say that data are MAR is to say that the missing-data mechanism is ignorable, i.e., whether or not there is data missing for some observation is completely dependent on a combination of random error and the observed values of other variables within that observation. If there is missing data on variables Y_{i2}, Y_{i3}, and Y_{i4}, then the mechanism for their missingness must be dependent solely on random error and Y_{i1} for the missing-data mechanism to be MAR. MAR data might be thought of as "conditionally missing completely at random"; the missing-data mechanism is nothing but noise after observed values of the predictors are controlled for.

Data are considered to be missing completely at random (MCAR) when the missing-data mechanism does not depend on the values of the observed values; that is, when

8

$$f(Y_{obs}, R | \Phi, \psi) = f(Y_{obs} | \Phi) f(R | \psi). \qquad \text{(MCAR data)} \qquad (2.4)$$

If neither of these conditions holds, then the observed values depend on Y_{mis}; this kind of data is referred to as NMAR (not missing at random).

No literature exists on whether meta-analytic data are generally MCAR, MAR, or NMAR in nature. However, it is easy to see how any of these types of missing data might arise in a group of studies. Suppose a sub-field in an area tends to find stronger correlations between two variables of interest than another sub-field that studies the same relationship in a different context. Because they are two different sub-fields, one will likely report variables the other does not, and vice-versa. The pattern of missingness on the predictors will be related to the outcome (the correlations), i.e., it will be MAR. MAR data will also arise if the missingness of any predictor (e.g., treatment length) is a function of the value of another predictor that is always observed (e.g., whether the program was private or public, or whether the paper the data was taken from was published in a peer-reviewed journal).

Alternatively, assume that the missingness of any one predictor is in part related to the values of any of the other predictors for which there is missing data. Or, assume that the outcome (e.g., correlations) is unreliably measured and there is a relationship between the outcome and the missing data pattern. In such a case, the missing data pattern will be NMAR.

2. Estimation Techniques for Datasets with Missing Data

There are numerous procedures one can use to estimate parameters in a general linear model (including a meta-analytic model) with missing data. By far the most common method is complete-case analysis, in which only those studies that have complete data are used, usually in an OLS or WLS regression. Other methods to be considered for application to a meta-analytic model are available-case analysis, unconditional-mean imputation, conditional-mean imputation (Buck's method), and maximum-likelihood estimation (MLE). I consider each of these in turn with regard to evidence of their estimators' bias in the face of data that are MCAR, MAR, and NMAR. Finally, I discuss the ramifications of this literature with regard to the specific nature of meta-analytic data and models.

Complete-Case Analysis

Complete-case (CC) analysis, also known as *listwise deletion,* is the simplest estimation technique, and the most common. As with most of the other estimation techniques discussed below, complete-case analyses give unbiased estimators of population correlation matrices, slopes, and random-effects variance terms when the data are MCAR. When the data are MAR, and the missing data are confined to the predictors (the latter is common in meta-analysis, though perhaps only because no methods have been developed to handle missing data on the outcomes), CC analyses give unbiased estimators of the slopes in the underlying regression model, but biased estimators of the population correlation matrix and means of the predictors (Glynn et. al., 1986; Little &

10

Rubin, 1987). This somewhat counterintuitive fact stems from the same statistical

argument that restriction of range on a predictor in a multiple regression causes a biased

(i.e., lower) estimate of the population correlation between predictor and outcome, but

does not affect the estimate of the slope for that predictor-outcome relationship.

While CC analyses give unbiased estimators of slopes even when the data are

MAR, their estimators are inefficient compared to alternative estimators. CC analyses risk

the loss of a significant fraction of one's sample of cases (or, in the meta-analytical

context, studies). This loss of information can cause a large drop in efficiency compared to

estimators from other estimation procedures (Kim & Curry, 1977; Little & Rubin, 1987;

Little & Raghunathan, 1999). Also, while CC estimators are unbiased when the data are

MAR, the same cannot be said when the data are not missing at random. Few studies have

investigated robustness of procedures with regard to data that are NMAR, however, so

the evidence cannot be considered to be strong in any direction. The little research that has

been done simulating NMAR data (e.g., Little, 1992; Little & Raghunathan, 1999)

suggests that CC analyses can lead to large biases relative to other estimation procedures,

as well as poorer confidence interval coverage. Any procedure gives biased results when

the data are NMAR if the estimation procedure is not specifically modeled to handle the

missing-data mechanism that is causing the data to be NMAR (Little & Rubin, 1992, Ch.

12); however, the research suggests CC estimators have noticeably worse biases than

estimators from other procedures.

Available-Case Analysis

There is good reason for distaste for the idea of throwing out otherwise good data, merely because it is incomplete. A natural idea is to try pairwise deletion of the data (as opposed to listwise). This kind of deletion leads to what is known as available-case (AC) analysis. However, AC analysis has a potential problem in that different sets of cases are used to estimate different means and covariances among a group of variables. Depending on how the means and covariances are estimated, the covariance matrix has the chance of not being positive definite. A lack of a positive definite covariance matrix makes multiple regression analysis difficult.

However, low correlations among variables minimize the risk of finding a positive indefinite matrix. Simulation studies have shown that when multiple regressions are based on population correlation matrices that have low to moderate values for the correlations, AC performs well compared to CC methods (Kim & Curry, 1977; Little, 1992). In fact, in Little's paper the AC analysis performs comparably to the maximum likelihood analysis. There is no evident bias for the AC estimators when the data are MCAR, a small bias for one predictor when the data are MAR, and the bias that arises when the data are NMAR is no greater than it is for the ML estimation. While the bias for the one predictor (the one that had missing values) cannot be ignored, it was not that much greater than the difference between the ML estimate and the correct value (.154 vs .113, with the AC analysis having a standard error of .364 and the ML analysis a standard error of .471). For the other predictors, AC analysis performed equally well or better than MLE, and considerably better than CC.

12

In Little's 1992 study, standard errors for the estimates of the AC slopes are 6-8% larger than the standard errors for the ML slopes. Little points out that the standard errors he reports, taken from a BMDP algorithm (Dixon, 1988) seem to have "no theoretical basis" and "appear too small". He suggests that correct estimators of standard errors require more complex formulas (e.g., see Van Praag, Dijkstra, & Van Veltzen, 1985).

There is one other perceived problem with AC analysis, and that is that it only performs this well when the intercorrelations among the variables are "low", as they were in Little's 1992 study. When intercorrelations are high, AC analyses decline in performance to where CC analyses are superior (Azen & Van Guilder, 1971; Haitovsky, 1968).The natural question to ask is, how low is "low" and how high is "high"? As it turns out, the correlations need not be very low at all for AC analyses to provide more efficient estimators than CC analyses. Kim and Curry (1977) point out that Haitovsky's 1968 paper compared AC and CC analyses using parameter values rarely found in social science data; all but two of the simulated multiple regressions had population multiple R^2s of greater than .7, and half had multiple R^2s of greater than .9! The two models that had lower R^2s (.596 and .158) were still unrepresentative of social science regression models as even these models had some correlations between the predictors that were greater than .8 and .9. Thus, the generalizability of Haitovsky's paper (and rarely is any other work cited regarding the weakness of AC estimation when intercorrelations are high) is limited at best.

A similar problem is found in Azen and Van Guilder (1971), in that most of the parameter values examined would rarely be found in a meta-analysis. Four conditions were

13

investigated with regard to the correlation structure of the predictors and outcome. Only the first, where $R^2 = .5$ and ρ (the correlation between the predictors) $=.25$ resembles anything that might be found in a meta-analysis; many would argue that an R^2 of .5, corresponding to a multiple R of .71, is unrealistically high. The other conditions were $R^2=.5$, $\rho=.75$, $R^2=.9$, $\rho=.25$, and $R^2=.9$, $\rho=.75$. Rarely do we see R^2s this high or correlations between predictors in a multiple regression this high anywhere in the social sciences.

Not surprisingly, Azen and Guilder found that when R^2 is high, available-case analysis should not be used. Similarly, they found that when the correlation structure among the predictors is strong (i.e., ρ is high), available-case analysis should not be used. They did find that in a case we are interested in, where $R^2=.5$ and $\rho=.25$, that "[Available case analysis] performs adequately when the . . . data is missing at random [MCAR] or in a related pattern [MAR]" (p. 54). When the data were NMAR ("truncated" in their language), they found that when $R^2=.5$ and $\rho=.25$, complete-case analysis performed somewhat better than did the maximum-likelihood estimation through the EM algorithm, which in turn performed slightly better than available-case analysis. This is a curious finding, as no where else in the literature has anyone found complete-case analysis to be superior to EM (or to available-case analysis when the correlations are low). Given that their study was based on only 50 replications, and given the differences between the three methods were slight for the condition of interest, these results must be interpreted with caution.

Kim and Curry's study simulated correlation matrices of five variables; the

intercorrelations varied between .322 and .596; they then investigated how well

CC methods estimated the population correlation and covariance matrices. AC

were superior in all cases, though sometimes the improvement was only modest

trivariate regression case where the proportion of missing values was uniform,

(1964) found that AC analyses give more efficient estimators than CC analyses

the correlation between the two independent variables is less than .58 -- and the

correlation between two independent variables is almost always less than .58 in

sciences, including study characteristics in a meta-analysis. Thus, while Kim and

(1977) are often cited as finding evidence that AC dominates CC analyses when

correlations are "modest" (Little & Rubin, 1987) or "small" (Rubin & Schafer,

fact is that the correlations need not be "small" or "modest" at all.

Unconditional Mean Imputation and Conditional Mean Imputation

Reports of the poor performance of unconditional mean imputation are

commonplace in the literature on missing data (e.g., Anderson et al., 1983; Littl

1992; Pigott, 1994). Variances and covariances are usually severely underestim

because one is imputing the mean for all missing cases. If the estimators are not

to take into account this underestimation, they will be both biased and inefficien

Adjustments for this underestimation lead to equations for the variances and co

equivalent to those in available-case analysis.

Imputation of conditional means, also known as Buck's method (Buck,

superior to imputation of unconditional means. It involves regressing the missin

15

onto the observed variables, and treating estimates of the missing data as "real" in the follow-up multiple regression analysis. This procedure tends to lead to underestimation of variances and covariances, just as unconditional mean imputation does, but the problem is less serious due to the regressions. In the simulations in Little (1992), Buck's method performed similarly to the AC analyses: it was less efficient than ML estimation, but about as biased. However, problems exist with the estimation of standard errors when the data have anything but a monotone pattern of missingness. Also, this procedure so closely resembles in operation ML estimation (which is essentially an iterative Buck's method) that it seems unreasonable to implement something as complicated as Buck's method without going one step further, i.e., maximum-likelihood estimation.

Maximum-Likelihood Estimation

The method of maximum-likelihood is, generally speaking, to choose as estimates of parameters those values of the parameters that maximize the likelihood of observing whatever data has been collected. Often, when the data are complete, ML estimators can be achieved by a straightforward derivation. Unfortunately, when data are missing, ML estimation is not so easy, as no closed form solutions of the log-likelihood equations exist. Many iterative methods have been proposed to obtain ML estimators even though no closed form solution exists, such as Fisher-scoring and the Newton-Raphson algorithm. However, as Little and Rubin (1987) point out, both of these methods require calculating the matrix of 2nd derivatives of the log-likelihood, which can be mathematically quite challenging. Two maximum-likelihood methods do not require this: multiple imputation

16

(Schafer, 1997a) and the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Little & Rubin, 1992). The formulas for these methods are similar (Schafer, 1997a); the chief advantage of multiple imputation over EM is that it allows for the calculation of standard errors. Most simulation research investigating the bias and efficiency of maximum-likelihood estimators that account for missing data has been done using the EM algorithm, however; the strongest multiple imputation work is very recent. To some extent, the method used to conduct a maximum-likelihood estimation is moot, as they are all based on the same maximum-likelihood equations and asymptotically should give the same results.

As with CC and AC estimation, ML estimation gives unbiased estimators of slopes when the data are either MCAR or MAR. There are some important distinctions between ML estimation and CC or AC estimation, however. When there is any non-trivial degree of missingness in the data, ML estimators dominate CC estimators with regard to efficiency, and this margin grows larger when the data are NMAR (Little, 1992; Little & Raghunathan, 1999). The difference between AC and ML estimation is less stark; it seems that when the intercorrelations among the predictors are low to moderate, AC estimators perform similarly to ML estimators (Little, 1992), though as mentioned above the AC estimators may be more biased and questions remain as to the correct standard errors for AC estimators. Unfortunately, all the evidence we have with regard to this comparison is from Little (1992), and it is dangerous to generalize from one study.

Besides Little (1992), only two studies (Little, 1988; Little & Raghunathan, 1999) contain findings from simulation research regarding how robust maximum likelihood

17

estimation is when the data are NMAR. In the first, Little tested through simulation how well two estimation techniques — one which assumed the errors were distributed as a multivariate t, and one which assumed normal errors — performed when the data were MCAR, MAR, and NMAR. His results indicate that when the data were NMAR, the ML estimation was about as good (i.e., very good) as when the data were MAR.

It is doubtful, however, that these results generalize to the meta-analysis case. In Little's simulations, the population correlation matrix was

$$\begin{bmatrix} 1 & .95 & .45 & -.77 \\ & 1 & .52 & -.58 \\ & & 1 & .08 \\ & & & 1 \end{bmatrix}$$

Variable 1 was always observed; variables 2, 3, and 4 were missing individually and in combinations across the data set. When the data were generated to be NMAR, the missing-data mechanism R was taken to be a function of the value of the 4th variable. Notice the weak relationship of variable 4 with variable 3 (only .0763) and the very, very strong correlation between variable 1 and variables 2 and 3 (respectively, .948 and .447). Variable 1 was never missing, meaning it could be used as a predictor of variables 2 and 3 in all cases within the EM algorithm. Unfortunately, the results of this study do little to answer the question of whether maximum likelihood methods are robust to NMAR.

In the study by Little and Raghunathan (1999), the authors investigate how well different types of EM estimation estimated slopes and a random effects term in a multi-level model with varying types of missing data. When their data were NMAR, confidence

18

interval coverages were generally very poor and the estimators had large biases.

3. Summary

Of the estimation techniques discussed, only two seem likely to give viable estimators for the meta-analysis model when significant amounts of data are missing: available-case estimation, and maximum-likelihood estimation. The differences between the two are twofold: AC analysis is methodologically simpler than MLE analysis, but the MLE analysis is more statistically rigorous. An AC analysis could theoretically be conducted on a spreadsheet, or through SPSS or SAS generated pairwise correlations combined with some pencil-and-paper work. A MLE analysis, on the other hand, would require specialized software. This distinction is an important one in that most meta-analyses are conducted by subject-matter specialists, who may or may not be statistically adept or expert at programming. Thus, while statistically the MLE analysis is likely to be superior in terms of bias and MSE, it is an open question as to whether it would give substantively different results from AC analysis.

CHAPTER III

MODELS AND ESTIMATION THEORY

1. The Meta-Analytic Model

Consider a meta-analysis of k independent effect magnitudes, T_i ($i = 1$ to k).

Typically,

$T_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + u_i + e_i$ (from $i = 1$ to k), where $X_{1i} \ldots$

X_{pi} are k x 1 matrices of data on the p predictors. If $X_i = \{X_{1i} \mid X_{2i} \mid \ldots \mid X_{pi}\}$ and $\beta =$

$\{\beta_0 \mid \beta_1 \mid \ldots \mid \beta_{p-1}\}'$, this may be restated as

$T_i = X_i\beta + u_i + e_i$, where $\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad$ (3.1)

$e_i \sim N(0, \sigma_i^2)$, and σ_i^2 is considered known[1],

$u_i \sim N(0, \tau)$, and

$X_i \sim N(\mu_X, \Sigma_X)$.

Furthermore, assume that for any given study i, data may be missing on any of the

predictors in X_i. Estimators of β and τ, and standard errors for these estimators, are

straightforward for the this model when no data are missing (see, e.g., Hedges & Olkin,

1985; Raudenbush, 1994). However, the presence of missing data makes matters complex.

A search of the literature found no available-case estimation procedure that would

be directly applicable to the meta-analysis equation in 3.1 (i.e., estimation with a fixed

---

[1] σ_i^2 is the fixed-effects variance of the errors. $\sigma_i^2 = V(e_i) = 1/(n_i - 3)$ in the case of Fisher's-zs, and $V(e_i) = 1/n_i^E + 1/n_i^C + \theta_i^2/2(n_i^E + n_i^C)$ for standardized mean differences (ds).

variance term for one variable and random-effects terms for all variables). Similarly, no maximum-likelihood procedure was found that, without some modification, could be used to estimate the model in 3.1. Some models are very similar; Bryk and Raudenbush (1992) give a model with both a known variance term and a random-effects variance to be used for meta-analysis. However, their estimation procedure does not take into account missing data on the predictors. Schafer (1997b) outlines a general multivariate model into which the meta-analysis model falls. It would treat all of the study characteristics as outcomes, and estimate both fixed-effects and random-effects variance-covariance matrices. However, Schafer's model assumes that the fixed-effects variance-covariance matrix needs to be estimated. This matrix is considered to be known in meta-analysis: the variance of the outcome is σ_j^2, and the variances of the predictors, and covariances between the Xs and the outcome are zero (as the covariances between the Xs and the outcome are best represented in the random-effects covariance matrix, not the fixed-effects covariance matrix). This makes Schafer's EM formulas impossible to use as they require an invertible fixed variance term. Finally, as already mentioned, the method given in Pigott (1992) accounts for missing data and was developed with the meta-analytic context in mind. However, it does not account for the presence of random-effects variation. Also, her model assumes that the variation in the predictors is proportional to the sample size of the study in question, which often will not be the case. For instance, if one study characteristic under investigation is the length of a treatment program, it would probably be wrong to expect variation in that characteristic to be smaller for studies with large numbers of subjects than with small numbers of subjects.

21

2. Estimation Procedures

Below I describe how I will estimate the model in 3.1 employing three different methods: complete-case estimation, available-case estimation, and maximum-likelihood estimation.

Complete-Case Estimation

The complete-case estimation follows the procedure described in Raudenbush (1994). The estimator of $\hat{\beta}$ is the standard WLS estimator $(X'V^{-1}X)^{-1}X'V^{-1}T$, where V is the diagonal $k \times k$ matrix of fixed effects variances. The WLS estimator of the population random-effects variance, τ, is found using the formulas given in Appendix A of Raudenbush (1994).

Available-Case Estimation

Little and Rubin (1992) outline three different methods to get pairwise covariance matrices from datasets with missing data. Simulation work led to the conclusion that in the majority of cases, the estimation procedure used to constrain correlations between the estimates to between -1 and 1 (Matthai, 1951) led to estimators with the greatest precision and fewest outliers. This method does not constrain the covariance matrix to be positive definite, however.

Define $Y = (Y_{ij})$ as an $k \times (p+1)$ matrix of k observations incompletely measured for $(p+1)$ variables ($i = 1$ to k, $j = 1$ to $p+1$). The matrix Y consists of one column for the outcome, T_i, and p columns for the p predictors. Introduce variable indexes v and w,

22

where v and w vary from 1 to $(p+1)$. Let a statistic's superscript (e.g., $^{(vw)}$ or $^{(v)}$) represent the variable(s) for which complete observations are necessary to calculate that statistic. The estimator of the covariance between two variables v and w (s^*_{vw}) is

$$s^*_{vw} = s^{(1)} \frac{\sqrt{s^{(v)}_{vv} s^{(w)}_{ww}}}{\sqrt{s^{(vw)}_{vv} s^{(vw)}_{ww}}}, \tag{3.2}$$

where

$$s^{(1)} = \sum_{(vw)} (y_{iv} - \bar{y}^{(vw)}_v)(y_{iw} - \bar{y}^{(vw)}_w) / (k^{(vw)} - 1). \tag{3.3}$$

To estimate the slopes and the intercept after the covariance matrix is estimated, I use the sweep operator defined in Beaton (1964). An excellent explanation of its use with regard to handling missing data appears in Chapter 6 of Little and Rubin (1992). It is straightforward to use each of the estimation procedures above to calculate pairwise covariance matrices. Weighted pairwise covariances might also be found; the formulas would change only in that each term in the equations is multiplied by the relative weight $w_i/\Sigma w_j$, where w_i is based on the inverse of the sampling error for study i. The weight w_i can be equal to zero based on the missingness of y_{iv}, y_{iw}, or both (depending on the formula). However, this addition severely complicates the procedure, especially for the non-statistician, given the many different weights that must be calculated for each variance and covariance. Thus, unweighted pairwise covariance matrices are estimated in the

method proposed[2]. Insofar as we are interested in estimating slopes, this procedure is a pairwise equivalent of OLS meta-analysis. There are two other key elements of a meta-analysis, however: an estimator of the random-effects variance (τ), and estimators of standard errors for the slopes.

The estimation of τ when data are complete is straightforward: we calculate the residual sums of squares, $\sum(T_i - X_i\hat{\beta})^2$, divide by $(k\text{-}p\text{-}1)$, and subtract from that the average study sampling error. While this method does not take into account the study weights, Raudenbush (1994) states that it yields an approximation that tends to be quite accurate. However, the residual sums of squares (RSS) cannot be directly calculated because of the missing Xs. We can bypass this difficulty by calculating the variance of the T_i and subtracting from it the variance explained by the model, $\hat{\beta}'\Sigma\hat{\beta}$. After a degrees-of-freedom adjustment which corresponds exactly to use of an adjusted R^2 in place of a normal R^2 to explain the variance explained in the sample, we arrive at

$$\sigma_T^2 = \hat{\beta}'\Sigma\hat{\beta} + \overline{\sigma}_e^2 + \tau, \text{ so } (\sigma_T^2 - \hat{\beta}'\Sigma\hat{\beta}) * k / (k - 4) \cong \overline{\sigma}_e^2 + \tau. \qquad (3.4)$$

The estimator of τ is found by subtracting the average study sampling error from the residual variance. A straightforward procedure to test H_0: $\tau = 0$ is to calculate what the value of the heterogeneity statistic, Q_E, would be given the estimator of τ. From Raudenbush (1994), we get $Q_E = ([(\hat{\tau}*\Sigma w_i) / k] + 1)*(k\text{-}p\text{-}2)$. If we were working with

[2] This begs the question of whether a weighted pairwise procedure might perform substantially better than an unweighted procedure. Both were used in the simulation study described in Chapter IV; while the weighted procedure occasionally gave more precise and less biased estimators than the unweighted procedure, the reverse was also true. In general, the difference between the two procedures in overall bias and MSE was slight.

complete data, Q_E would be distributed as a chi-square with k-p-2 degrees of freedom when the null hypothesis $\tau = 0$ is true. However, there is not complete data on the p predictors. An intuitive adjustment to the degrees of freedom, used in a related context by Su (1988), is to multiply p by the fraction of missing data across the p predictors.

Maximum-Likelihood Estimation

Pigott (1992) showed that through a weighted EM estimation one could conduct a fixed-effects meta-analysis with missing data on the predictors. However, she noted that her algorithm treats all variables in the model, both predictors and outcome, as if the precision with which they are measured is proportional to the sample size of the study. As she points out (p.53), this assumption may not be a valid one for data from a research synthesis. Below, I present a method for doing a weighted EM estimation that assumes that only the outcome is measured with a precision proportional to the size of the study. The predictors are assumed to be measured with equal precision across all studies and with no measurement error, regardless of the size of the study. My method also allows for the presence of a random-effects error term.

The goal of this section is to obtain estimators of $X_{M,i}$, β, Σ_X, μ_X, u_i, and τ through the EM algorithm. After finding sufficient statistics for the parameters that are to be treated as fixed, the distributions of the parameters to be treated as random (conditioning on the fixed parameters and the observed data) are derived, and estimators of the parameters treated as random are found. This is the "E-step". During the "M-step", these estimates are used in the sufficient statistics to obtain new estimates of the fixed

25

parameters. The new estimates of the fixed parameters lead to new estimates of the random parameters, and the process iterates until convergence.

The E-Step

Dropping subscripts, for each study i, the joint distribution of T, X_O, X_M, and u is as follows:

$$
\begin{bmatrix} T \\ X_O \\ X_M \\ u \end{bmatrix} \sim N \left[\begin{pmatrix} \beta\mu_X \\ \mu_O \\ \mu_M \\ \beta_0 \end{pmatrix}, \begin{pmatrix} \beta'\Sigma_X\beta + \sigma_i^2 + \tau & \beta_O'\Sigma_O + \beta_M'\Sigma_{M,O} & \beta_M'\Sigma_M + \beta_O'\Sigma_{O,M} & \tau \\ \Sigma_O\beta_O + \Sigma_{O,M}\beta_M & \Sigma_O & \Sigma_{O,M} & 0 \\ \Sigma_M\beta_M + \Sigma_{M,O}\beta_O & \Sigma_{M,O} & \Sigma_M & 0 \\ \tau & 0 & 0 & \tau \end{pmatrix} \right].
$$

(3.5)

where

$$
X = \begin{bmatrix} X_O \\ X_M \end{bmatrix}, \quad \Sigma_X = \begin{bmatrix} \Sigma_O & \Sigma_{O,M} \\ \Sigma_{M,O} & \Sigma_M \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_O \\ \beta_M \end{bmatrix} \quad\quad (3.6)
$$

For study i, X_O consists of the observed Xs and X_M consists of the unobserved Xs. Similarly, β_O consists of the population slopes for the observed Xs and β_M consists of the population slopes for the missing Xs. The column of 1's typically included in the matrix of predictor variables is excluded, as β_0, the intercept, is defined as the mean of the random effects.

In this multivariate representation of the model for the ith study, only X_M and u

26

are treated as random. All other parameters, such as Σ_X, μ_X, and the slopes for the study

characteristics, β, are treated as fixed. It is standard when doing MLE/EM estimation to

assume that the variance-covariance matrices are fixed (Bryk & Raudenbush, 1992; Little

& Rubin, 1987; Pigott, 1992). This assumption is what differentiates the empirical Bayes

solution from an exact Bayes solution, which can be extraordinarily computationally

complex (Bryk & Raudenbush, 1992). When the assumption that β is fixed is made, the

type of estimation being done is often referred to as full maximum likelihood (MLF)

estimation. When the assumption is not made, and a non-informative prior is specified for

β instead, it is referred to as restricted maximum likelihood (MLR) estimation. The

difference between these two, derivation-wise, is not trivial. MLF estimation is easier than

MLR, as fewer parameters are treated as random. Common wisdom is that in the end, the

two usually give similar results except that MLR gives a degrees-of-freedom correction

for the estimators.

There is more to the difference between MLF and MLR than a degrees-of-freedom

correction, however. In a meta-analysis working with complete data, MLR estimation of β

weights studies by $1/(\tau+\sigma_i^2)$; see, for instance, the V-known model in Bryk and

Raudenbush (1992) or Shadish and Haddock (1994). However, MLF estimation of β

weights studies by $1/\sigma_i^2$. The reason for this stems from the lack of a non-informative prior

for β in 3.5. Equation 3.5 treats β as a fixed effect; were we able to treat it as a random

effect (and thus use the non-informative prior to describe its variance), the weighting

would be different. The derivation would be made much more complicated with this step,

however, as it would mean both the Xs and β are being treated as random.

27

Another difference is mentioned in Bryk and Raudenbush (1992); when the number of level-2 units (here, studies) is small, MLF estimators of τ can also be too small. This problem is exacerbated the more fixed effects the model contains.

In Equation 3.5, we assume that for study i, T and X_O are fully observed random variables, while X_M and u are unobserved random variables, and that we have estimates of the fixed parameters τ, Σ_X, β, and μ_X. Standard multivariate normal distribution theory (e.g., Morrison, 1967, p.88) is used to get $E(X_M, u \mid \tau, \Sigma_X, \beta, \mu_X)$; the estimators of X_M and u conditioned on T and X_O. Call these estimators X_M^* and u^*, specifically,

$$
\begin{pmatrix} X_M^* \\ u^* \end{pmatrix} = \begin{pmatrix} \mu_M \\ \beta_0 \end{pmatrix} +
$$

$$
\begin{pmatrix} \Sigma_M \beta_M + \Sigma_{M,O}\beta_{M,O} & \Sigma_{M,O} \\ \tau & 0 \end{pmatrix} \begin{pmatrix} \beta'\Sigma_X\beta + \sigma_i^2 + \tau & \beta_O'\Sigma_O + \beta_M'\Sigma_{M,O} \\ \Sigma_O\beta_0 + \Sigma_{O,M}\beta_M & \Sigma_O \end{pmatrix}^{-1} \begin{pmatrix} T - \beta_0' - \beta'\mu_X \\ X_O - \mu_O \end{pmatrix} \tag{3.7}
$$

We also need $\mathrm{Var}(X_M, u \mid \tau, \Sigma_X, \beta, \mu_X)$. We call this matrix D^*, which is

$$
D^* = \begin{pmatrix} D_X^* & D_{X,u}^* \\ D_{u,X}^* & D_u^* \end{pmatrix} = \begin{pmatrix} \Sigma_M & 0 \\ 0 & \tau \end{pmatrix} -
$$

$$
\begin{pmatrix} \Sigma_M \beta_M + \Sigma_{M,O}\beta_{M,O} & \Sigma_{M,O} \\ \tau & 0 \end{pmatrix} \begin{pmatrix} \beta'\Sigma_X\beta + \sigma_i^2 + \tau & \beta_O'\Sigma_O \\ \Sigma_O\beta_0 & \Sigma_O \end{pmatrix}^{-1} \begin{pmatrix} \beta_M'\Sigma_M + \beta_O'\Sigma_{O,M} & \tau \\ \Sigma_{O,M} & 0 \end{pmatrix}. \tag{3.8}
$$

These expressions become necessary in the M-step, where the sufficient statistics for the fixed parameters are used to estimate the fixed parameters.

<u>Sufficient Statistics for β, τ, Σ_X, μ_X</u>

Maximum-likelihood estimation in EM proceeds by assuming that we have complete data for all of the parameters that we are treating as random. As we are treating T, X_O, X_M, and u as random, we need the complete-data joint maximum likelihood of these four parameters, given all other parameters. Although, for any given study, there is an X_O and X_M, if we are assuming we have complete data we can (temporarily) unpartition them and refer to the predictors collectively as X.

The key to the derivation is to take advantage of the identity $f(m,n) = g(n)h(m|n)$.

Thus, we have

$$f(T, X, u \mid \tau, \Sigma_X, \beta, \mu_X) = g_1(T \mid X, u, \tau, \Sigma_X, \beta, \mu_X)h(X, u \mid \tau, \Sigma_X, \beta, \mu_X)$$

$$= g_1(T \mid X, u, \tau, \Sigma_X, \beta, \mu_X)g_2(X \mid u, \tau, \Sigma_X, \beta, \mu_X)g_3(u \mid \tau, \Sigma_X, \beta, \mu_X)$$

$$= g_1(T \mid X, u, \tau, \Sigma_X, \beta, \mu_X)g_2(X \mid \tau, \Sigma_X, \beta, \mu_X)g_3(u \mid \tau, \Sigma_X, \beta, \mu_X),$$

given the independence of X and u; across the k studies, the values of the predictors and the values of the random-effects error terms are unrelated.

This allows us to break the likelihood into three tractable pieces:

$$L[f(T, X, u \mid \tau, \Sigma_X, \beta, \mu_X)] = L[f(T \mid X, u, \tau, \Sigma_X, \beta, \mu_X)]$$

$$* L[g(X \mid \tau, \Sigma_X, \beta, \mu_X)] * L[h(u \mid \tau, \Sigma_X, \beta, \mu_X)],$$

$$L = \prod_{i=1}^{k} f(T|X,\beta,\tau,\Sigma_{X,\mu_X}) \prod_{i=1}^{k} g(X|\beta,\tau,\Sigma_{X,\mu_X}) \prod_{i=1}^{k} h(u|\beta,\tau,\Sigma_{X,\mu_X})$$

$$= \prod_{i=1}^{k} \frac{1}{\sqrt{(2\pi)\sigma_i^2}} \exp\left[\frac{-\frac{1}{2}(T_i - X_i\beta - u_i)'(T_i - X_i\beta - u_i)}{\sigma_i^2}\right] *$$

$$\prod_{i=1}^{k} \frac{1}{\sqrt{(2\pi)^P \Sigma_X}} \exp\left[-\frac{1}{2}(X_i - \mu_X)' \Sigma_X^{-1}(X_i - \mu_X)\right] * \qquad (3.9)$$

$$\prod_{i=1}^{k} \frac{1}{\sqrt{(2\pi)\tau}} \exp\left[\frac{-\frac{1}{2}u_i' u_i}{\tau}\right].$$

This leads to the log-likelihood

$$\log L = -\frac{k}{2}(p+2)(\log 2\pi) - \frac{1}{2}\sum_{i=1}^{k}\log \sigma_i^2 - \frac{k}{2}\log|\Sigma_X| - \frac{k}{2}\log \tau -$$

$$= \sum_{i=1}^{k}\left[\frac{\frac{1}{2}(T_i - X_i\beta - u_i)'(T_i - X_i\beta - u_i)}{\sigma_i^2}\right] - \qquad (3.10)$$

$$\sum_{i=1}^{k}\left[\frac{1}{2}(X_i - \mu_X)' \Sigma_X^{-1}(X_i - \mu_X)\right] - \sum_{i=1}^{k}\left[\frac{1}{2}u_i' \tau^{-1}u\right].$$

All that remains is to differentiate with respect to the fixed parameters τ, Σ_X, β, and μ_X and find the sufficient statistics for these parameters. After the sufficient statistics have been derived, the formulas for the expected values and variance-covariance matrix of X_M^* and u^* are used to complete the EM process. The formulas are

$$\frac{d \log L}{d\tau} = -\frac{k}{2}\tau^{-1} + \frac{\tau^{-2}}{2}\sum_{i=1}^{k} u_i' u_i,$$

$$\hat{\tau} = \frac{1}{k}\sum_{i=1}^{k} u_i' u_i.$$

$$\frac{d \log L}{d\mu_X} = \frac{1}{2}\sum_{i=1}^{k}\left[(2X_i\Sigma_X^{-1}\vec{1}) - (2\mu_X\Sigma_X^{-1}\vec{1})\right],$$

where $\vec{1}$ is a $px1$ vector of 1's,

$$\hat{\mu}_X = \frac{1}{k}\sum_{i=1}^{k} X_i.$$

$$\hat{\Sigma}_X = \frac{1}{k}\sum_{i=1}^{k}(X_i - \hat{\mu}_X)'(X_i - \hat{\mu}_X).$$

$$\frac{d \log L}{d\beta} = \sum_{i=1}^{k}\frac{1}{\sigma_i^2}(T_i - X_i\beta - u_i)'(-X_i),$$

$$\sum_{i=1}^{k}\frac{1}{\sigma_i^2}X_i'X_i\beta = \sum_{i=1}^{k}\frac{1}{\sigma_i^2}X_i'(T_i - u_i),$$

$$\hat{\beta} = \left[\sum_{i=1}^{k}\frac{1}{\sigma_i^2}X_i'X_i\right]^{-1}\sum_{i=1}^{k}\frac{1}{\sigma_i^2}X_i'(T_i - u_i).$$

The log-likelihood allows us to use results from Tatsuoka (1988, p. 410) to g

estimator of Σ_X. The estimator of β has weights that do not involve τ because the u_i

have been partialled out, as shown in Searle et al. (1992), on page 297.

31

The sufficient statistics for the maximum likelihood estimators are

$$\sum_{i=1}^{k} X_i, \ \sum_{i=1}^{k} X_i' X_i, \ \sum_{i=1}^{k} \frac{1}{\sigma_i^2} X_i' X_i, \ \sum_{i=1}^{k} \frac{1}{\sigma_i^2} X_i' T_i, \ \sum_{i=1}^{k} \frac{1}{\sigma_i^2} X_i' u_i, \ \sum_{i=1}^{k} u_i' u_i. \qquad \textbf{(3.15)}$$

Note that we need both weighted and unweighted estimators of the sums of cross-products of the X_i.

The M-Step

The next stage of the EM algorithm is to calculate the expected value of the sufficient statistics after conditioning on the observed data. Doing so will allow us to estimate the parameters that we are treating as fixed, thus updating their estimates from whatever was used at the previous iteration. Below I index the p columns of X_i by subscripts r and s, and $P^{(t)}$ represents the estimates of the parameters that we are treating as random, i.e., $X_{M,i}^*$ and u_i^*, at iteration t. These parameters are shown in equation 3.7.

The expected value of X_{ir} (i.e., the expected value of predictor r, $r = 1 \ldots p$, in study i), given the observed data and our estimates is straightforward, specifically

$$E(X_{ir}|T_i, X_{O,i}, P^{(t)}) = \left\{ \begin{array}{l} X_{ir} \ \textit{if} \ X_{ir} \ \textit{is observed} \\ X_{ir}^{*(t)} \ \textit{if} \ X_{ir} \ \textit{is missing} \end{array} \right\}. \qquad \textbf{(3.16)}$$

The same is true for the expected value of $X_{ir}T_i$, which is

$$E(X_{ir}T_i \mid T_i, X_{O,i}, P^{(t)}) = \begin{cases} X_{ir}T_i & \text{if } X_{ir} \text{ is observed} \\ X_{ir}^{*(t)}T_i & \text{if } X_{ir} \text{ is missing} \end{cases} \qquad (3.17)$$

The expressions for the others are more complex because they include cross-products between "real" data and imputed data; that is,

$$E(X_{ir}X_{is} \mid T_i, X_{O,i}, P^{(t)}) =$$

$$\begin{cases} X_{ir}X_{is} & \text{if } X_{ir}, X_{is} \text{ are observed} \\ X_{ir}X_{is}^{*(t)} & \text{if } X_{ir} \text{ is observed, } X_{is} \text{ is missing} \\ X_{ir}^{*(t)}X_{is}^{*(t)} + Cov(X_{ir}^{*(t)}, X_{is}^{*(t)}) & \text{if both } X_{ir} \text{ and } X_{is} \text{ are missing} \end{cases} . \qquad (3.18)$$

The last expression arises from the identity $E(X_r^* X_s^*) = E(X_r^*)E(X_s^*) + Cov(X_r^* X_s^*)$. Here,

$$Cov(X_{ir}^{*(t)}, X_{is}^{*(t)}) = D_{X\,i,rs}^* \qquad (3.19)$$

Thus, we can find the necessary covariance by consulting the r'th row and s'th column of our solution for D_X^* in (3.8).

The solution for the weighted estimator of the sums of cross-products of the Xs is similar to the unweighted version; all that is added in are the weights, which are proportional to the inverse of the (known) values of σ_i^2. We have

$$E(w_i X_{ir} X_{is} | T_i, X_{O,i}, P^{(t)}) =$$

$$
\begin{cases}
w_i X_{ir} X_{is} & \text{if } X_{ir}, X_{is} \text{ are observed} \\
w_i X_{ir} X_{is}^{*(t)} & \text{if } X_{ir} \text{ is observed, } X_{is} \text{ is missing} \\
w_i X_{ir}^{*(t)} X_{is}^{*(t)} + w_i Cov(X_{ir}^{*(t)}, X_{is}^{*(t)}) & \text{if both } X_{ir} \text{ and } X_{is} \text{ are missing}
\end{cases}. \tag{3.20}
$$

A similar argument gives the best estimator of the sum of squares of the u_i's that is required in order to estimate τ:

$$E(u_i u_i | T_i, X_{O,i}, P^{(t)}) = u_i^{(t)} u_i^{(t)} + Var(u_i^{(t)}) = u_i^* u_i^* + D_{u,i}^*. \tag{3.21}$$

The estimator of the covariance between our estimate of X_i and u_i is in $D_{u,X}^*$ in (3.8). If the data were complete, the expected value of the product of X_{ir} and u_i would be zero. Because some X_{ir} must be estimated from the same data that is leading to estimates of the u_i, the product does not need to be exactly zero, especially in early iterations.

Finally, we have

$$E(w_i X_{ir} u_i | T_i, X_{O,i}, P^{(t)}) =$$

$$
\begin{cases}
w_i X_{ir} u_i^{(t)} & \text{if } X_{ir} \text{ is observed} \\
w_i X_{ir}^{*(t)} u_i^{*(t)} + w_i Cov(X_{ir}^{*(t)}, u_i^{*(t)}) & \text{if } X_{ir} \text{ is missing}
\end{cases}. \tag{3.22}
$$

From the above expected values we can get updated estimates of each of the parameters that are considered fixed: τ, Σ_X, β, and μ_X. These updated estimates allow us to calculate estimates $X_M^{*(t+1)}$, $D^{*(t+1)}$, and $u^{*(t+1)}$, and the process continues until convergence. Following Dempster et al. (1981), the likelihood function to be monitored

for convergence is $f(Y, X_O \mid \tau, \Sigma_X, \beta, \mu_X)$, and we can derive the likelihood function as follows:

$$f(T, X_O \mid \beta, \Sigma_X, \mu_X, \tau) = \frac{g(T, X_O, X_M, u \mid \beta, \Sigma_X, \mu_X, \tau)}{h(X_M, u \mid T, X_O, \beta, \Sigma_X, \mu_X, \tau)}, \qquad (3.23)$$

where $g(\ldots)$ was derived as in (3.9), and $h(\ldots)$ is the conditional distribution of the missing parameters treated as random given the observed random variables and the fixed parameters. The conditional distribution $h(\ldots)$ is normal with the mean given in (3.7) and variance-covariance matrix given in (3.8).

Simplification of the likelihood function relies on recognizing that the likelihood is true regardless of the values of X_M and u. Thus, suppose $X_M = X_M^*$ and $u = u^*$. This simplifies the denominator to a great extent as the terms in the exponent equal 0. After taking the log we are left with an expression similar to that in (3.10) but with a term representing the conditional variance-covariance matrix of X_M^* and u^* (namely, $k \log |D^*|$).

$$\log L(convergence) \propto -\frac{k}{2} \log |\hat{\Sigma}_X| - \frac{k}{2} \log \hat{\tau} + \frac{k}{2} \log |D^*| -$$

$$\sum_{i=1}^{k} \left[\frac{\frac{1}{2} \left(T_i - X_i^* \beta - u_i^* \right)' \left(T_i - X_i^* \beta - u_i^* \right)}{\sigma_i^2} \right] - \qquad (3.24)$$

$$\sum_{i=1}^{k} \left[\frac{1}{2} \left(X_i^* - \hat{\mu}_X \right)' \hat{\Sigma}_X^{-1} \left(X_i^* - \hat{\mu}_X \right) \right] - \sum_{i=1}^{k} \left[\frac{1}{2} u_i^{*'} \hat{\tau}^{-1} u_i^* \right].$$

The value needed for convergence is somewhat arbitrary, and depends on weighing the importance of accuracy against the reality that computer running time is limited. Simulations showed that a criteria of .000001% change between iteration i and $i+1$ resulted in estimators of β and τ that were within .0005 of the estimates found using the far more strict criteria of .000000001%, and often the estimates were closer.

MLE Standard Errors

Standard errors can be generated after maximum-likelihood EM estimation in several different ways. Little and Rubin (1987) explain how to get asymptotic standard errors for the slopes from the inverse of the observed or expected information matrix. These are two different methods, and neither the observed nor the expected information matrix is a natural output of the EM algorithm. Pigott (1992) states that the procedure for calculating the information matrix depends on the specific model being estimated. Other suggested procedures include bootstrapping (Little, 1988; Su, 1988), the SEM algorithm (Meng & Rubin, 1987), or use of an approximation formula (Beale & Little, 1975). Also, as noted above, the multiple imputation technique given in Schafer (1997a) allows for calculation of standard errors. After doing extensive simulation work, Su (1988) demonstrated that standard errors based on the observed information matrix are slightly superior to those found through the bootstrap method or the expected information matrix, but that the bootstrap method was less affected by model misspecification than the other variance estimators. He recommends use of bootstrapping standard errors in general, especially when robustness to model assumptions is of concern.

Given the relative ease of generating bootstrapped standard errors, and Su's positive review of the technique, the MLE program written to conduct meta-analyses allows for the calculation of bootstrapped standard errors for the slopes. The method follows Su (1988) generation of conditional bootstrapped (CBOOT) standard errors.

His method conditions on the observed missing data pattern (also known as the response pattern), R. The cases within each pattern are treated as independently and identically distributed (i.i.d.) random variables from the conditional distribution of the predictors and outcome (Y) given the observed pattern. The procedure of estimating the covariance matrix is as follows:

1. Let M be the total number of observed patterns and k_m be the number of cases in pattern m, $m = 1$ to M. For data in pattern m, draw a "bootstrap subsample" of k_m studies, with replacement, from those studies with pattern m.

2. Combine the M bootstrap subsamples into a "bootstrap sample" and calculate estimates of β, denoted β^*.

3. Independently repeat steps 1 and 2 B times, obtaining bootstrap replications $\beta^{*1} \ldots \beta^{*B}$.

4. Calculate the variances and covariances of the bootstrap replications:

$$C\hat{o}v(\hat{\beta}) = \frac{1}{B-1} \sum_{b=1}^{B} (\beta^{*b} - \bar{\beta}^*)(\beta^{*b} - \bar{\beta}^*). \tag{3.24}$$

The variances allow for the calculation of confidence intervals for the estimates of the slopes.

CHAPTER IV

SIMULATION STUDY METHODOLOGY

In this chapter, I outline a Monte Carlo study to evaluate the proposed maximum-likelihood and available-case estimators of the meta-analytic parameters in (3.1). I begin by describing the ranges of hyperparameter values for the simulations, and explaining how these ranges of values were determined. (The term "hyperparameter" is used to describe the variables varied in the simulations to differentiate them from the parameters that are to be estimated in the meta-analytic model.) I also describe how the data were generated using the values of the hyperparameters. The chapter concludes with a description of the criteria for measuring and comparing the performance of the complete-case, available-case, and maximum-likelihood estimators of β and τ, as well as for testing the bootstrapped standard errors for the maximum-likelihood method.

1. Hyperparameter Choices

I consider seven hyperparameters for variation within the simulation study: average study sample size (avg n_i), number of studies in the meta-analysis (k), random-effects variation (τ), study correlations between predictor and outcome and between predictors, variation in outcome explained by predictors, the incidence of missing data, and the type of missing data (MCAR vs. MAR vs. NMAR). Hyperparameters that do not change include the number of predictors (three) and the type of outcome (standardized mean differences).

38

The Outcome

The outcome to be simulated is the standardized mean difference, calculated using simulated raw data. Standardized mean differences are distributed as non-central t-statistics. This outcome was chosen instead of Fisher's z (the asymptotically normal transformation of the correlation coefficient) to see whether the maximum-likelihood method could handle the heavier tails of the non-central t-distribution.

Average Study Sample Size

Past simulations of meta-analytic data (Becker, 1985; Chang, 1992; Fahrbach, 1995) investigated study sample sizes ranging from 20 to 250. A survey of meta-analyses in *Psychological Bulletin* from 1995 to 1999 showed that average study sample sizes tend to range between 60 and 500, although there are exceptions in particular cases (e.g., if the primary sampling methodology in the field was mass community phone or mail surveys, or, conversely, if the primary study design was the case study). Based on this literature, two average study sample sizes were chosen for simulation: 80 and 400. The sampling errors for effect sizes generated using these study sample sizes are roughly equivalent to the sampling errors that would be generated if Fisher's zs were simulated for ns of 40 and 200. The former is near the lower end of what is considered an "acceptable" study sample size in the social sciences (Gay, 1992), and the latter is generally considered to be moderate to large.

There remains the question of how much the study sample sizes should vary. Optimally, there are three natural choices: none (equal study sample sizes), moderate, and

high (e.g., where the imbalance is such that the largest study's sample size may have half of the total sample in the meta-analysis). Because it is of interest to determine how well these estimation procedures work with weighted data, I assign study sample sizes in such a way that there is a high degree of imbalance across the k studies. The survey of recent meta-analyses in *Psychological Bulletin* found that most meta-analyses had large variation in study sample sizes, and that up to 60% of the total study sample size (N) could be within the largest one-fifth of the studies. I let one-fifth of the studies contain 50% of N, and the other four-fifths of the studies contain the other 50% of N.

Number of Studies per Meta-Analysis (k)

The simulation studies of Becker (1985) and Chang (1992) use values of k ranging between 2 and 50. The previously mentioned survey of meta-analyses published over the last five years in *Psychological Bulletin* showed meta-analyses ranging in size from 19 studies to over 300; most meta-analyses had at least 35 studies, and almost half had over 100.

An important factor in determining the values of k for this simulation study is the presence of missing-data mechanisms. Such mechanisms cannot be fairly represented in a meta-analysis with a very small number of studies. In addition, it is rare (and inadvisable) to examine moderator variable effects with small numbers of studies due to the amount of imprecision expected. Just as one would not recommend conducting a multiple-regression with an n of 20, one cannot recommend conducting a moderator meta-analysis with a k of 20.

Given these facts and the review of the *Psychological Bulletin* meta-analyses, values for k of 40 and 100 were selected.

Random-Effects Variation (τ)

The values for τ were 0, .005, and .02. The two nonzero values roughly correspond to 95% bands for ranges of the population effect size of ±.14, and ±.28, which translate roughly to small variation and large variation in the outcome after sampling error is accounted for. Cohen (1988) states that a small effect size is about .2, a moderate one, about .5, and a large one, about .8. Thus, values of τ much lower than .005 are often not substantively interesting (τ = .005 indicates that almost all studies are within .15 of the mean, and an effect size difference of .15 is considered small).

Population Correlation Matrix among Predictors and Outcomes

The slopes were generated by starting with the following set of correlations among the outcome (represented by the first row/column) and the predictors (represented by the 2nd, 3rd, and 4th rows/columns). The figures presented are rounded to two decimal places.

$$
\begin{bmatrix}
1 & .50 & .62 & .75 \\
.50 & 1 & .10 & .10 \\
.62 & .10 & 1 & .10 \\
.75 & .10 & .10 & 1
\end{bmatrix}
\qquad
\begin{bmatrix}
1 & .67 & .80 & .93 \\
.67 & 1 & .40 & .50 \\
.80 & .10 & 1 & .60 \\
.93 & .50 & .60 & 1
\end{bmatrix}
$$

In each matrix the underlying R^2 is 1.00, corresponding to a fixed effects model where the effect is fully explained by the Xs, and the predictors have varying strengths of relationships with the outcome. The primary difference is that in the first matrix, the predictors are weakly correlated, while they are strongly correlated in the second. Each of these correlation matrices were used to compute standardized betas to be used in the data generation process that is described below. The first matrix generates $\beta_1 = .38$, $\beta_2 = .52$, and $\beta_3 = .66$. The second matrix generates $\beta_1 = .22$, $\beta_2 = .34$, and $\beta_3 = .62$.

The underlying R^2 for the relationship between the predictors and outcome is unity in these matrices, but the sample R^2 will be lower due to two previously mentioned hyperparameters: sample size (which causes sampling variation), and τ (which causes random-effects variation). The sample R^2 will decrease an amount depending on the size of these two types of variation relative to that caused by the next hyperparameter.

Variation in Outcome Caused by Predictor Variables (V_{mod})

Correlation matrices alone cannot prescribe a relationship between the predictors and the outcome; the total amount of variation in the outcome caused by the predictors must also be set. This amount is defined by the hyperparameter, V_{mod}; it allows R^2 to vary across simulations while keeping βs, sampling error, and random-effects error constant. V_{mod} is used to scale the variance of the predictors. A description of how V_{mod} was used follows in section 2.

Values of V_{mod} of .006 and .03 were chosen, for reasons similar to those for choosing the values of τ. The first value corresponded to a small effect for the three

42

predictors, while the second value corresponded to a large effect. The underlying sample R^2s for $V_{mod} = .006$ ranged from around .06 (for high sampling error and large τ) to about .30 (for low sampling error and zero τ). $V_{mod} = .03$ led to sample R^2s ranging from .25 to .70.

Incidence of Missing Data

All simulated meta-analyses were of models with three predictors. While simulation with more predictors would be preferable, EM estimation is computer-intensive and there is a geometric relationship between the number of predictors in the model and the time it takes to estimate the parameters in that model. Past simulation research of MCAR and MAR data (e.g., Little, 1992; Su, 1988) provides little theoretical basis for choice of a specific missing-data mechanism or fraction of missing data on any given predictor. If anything, there is a focus on keeping things simple: Little (1992) simulates four predictors, but only the first has missing data (about 50%). Su (1988) has three predictors, and assumes the 2nd has 25% missing data and the third, 50%. Because no studies have described missing data patterns in meta-analytic datasets, there is little to rely on but these efforts, and good judgement.

Two patterns of missing data were generated. In both, the first predictor is always observed, while the second and third are sometimes missing. In the first pattern, 25% of the studies are missing information on both predictors, 25% of the studies are missing data on the 2nd predictor, 25% of the studies are missing data on the 3rd predictor, and 25% of the studies are complete. This results in both the 2nd and 3rd predictors each being missing

43

50% of the time. In the second pattern, 10% of the studies are missing data on both predictors, 20% of the studies are missing data on the 2nd predictor, 20% of the studies are missing data on the 3rd predictor, and 50% of the studies are complete. This results in both the 2nd and 3rd predictors each being missing 30% of the time. Appendix 1 contains details regarding these patterns and how they relate to variation in study sample size and the number of studies per meta-analysis. For simplicity's sake, these two conditions will be referred to based on the proportion of data that would be missing in a complete-case analysis, i.e., "75% Incidence of Missing Data" and "50% Incidence of Missing Data".

Types of Missing Data

As noted previously, there are three different types of missing-data mechanisms: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Within each missing data pattern, different percentages of data can be missing on each variable, and different strengths of relationships may hold between any given variable and the missing-data mechanism. For instance, assume that there are five predictors of interest, and only the first three are completely observed. For MAR data, missing data on the 4th and 5th variables might be a direct function of the value of the 1st variable; or it might be strongly related to the value of the 2nd variable; or it might be weakly related to a function of both the 1st and 2nd variable.

Generation of the MCAR pattern of missing data is straightforward. For instance, for the first incidence-of-missing-data pattern mentioned in the last section (75% Missing Data), the first 25% of the studies are treated as complete, the next 50% are missing

values, alternately, on the 2nd or the 3rd predictors, and in the last 25% both the 2nd and

3rd predictors' values are deleted. The variation in sample sizes is taken into account so

that after study deletion, no pattern has more large studies than any other pattern (e.g., it

would never be the case that most of the large studies are the studies in which data are

missing on the 2nd predictor, or the studies with complete data, etc.).

One type of MAR data was generated. The construction of the missing-data

pattern begins by generating a random normal deviate (r.n.d.) correlated .8 with the only

completely observed predictor, X_1. Each study thus has its own r.n.d. For any given meta-

analysis's missing data pattern and number of studies, x% of the studies will have

complete data, y% will have data missing on either the 2nd or 3rd predictors, and z% will

have data missing on two variables (x+y+z=100). For instance, for the first incidence

pattern described above, x% = 25%, y% = 50%, and z% = 25%. The studies with the top

x% values of the r.n.d. are treated as complete: no variables are unobserved. The studies

with the middle y% values of the r.n.d. are treated as partially observed; only one variable

(the 2nd or 3rd, chosen at random) is treated as missing[3]. In the studies with the bottom

z% values of the r.n.d., both the 2nd and 3rd predictors are treated as missing.

This process results in a correlation of approximately .50 between the value of the

completely observed predictor used to generate the selection bias, and a dummy variable

indicating whether or not for that study each predictor (2nd or 3rd) is observed. This

correlation is lower than the .80 value used to create the r.n.d. because "missingness" is

---

[3]   Within any given meta-analysis, this process is controlled such that the 2nd and 3rd
        predictors are each treated as missing half the time.

45

dichotomous, limited to 0's and 1's.

Two types of NMAR data were generated in which the missing-data mechanism is dependent on the true value of the 2nd (incompletely observed) predictor. Data generated using this type of mechanism will be collectively referred to as "predictor-NMAR data", or p-NMAR data (to distinguish them from another type of NMAR data described below). The process to generate these data is similar to that described above for the MAR data. One type assumes a strong relationship between the true values of the predictor and the chance that the 2nd or 3rd predictor will be missing; a correlation of .8 is used to create the r.n.d. These data will be referred to as sp-NMAR data, as there is a strong relationship between the predictor and the missing-data mechanism. The other type assumes a moderate relationship between the two variables; the correlation used to generate the r.n.d. is only .4. These data are referred to as mp-NMAR data.

One type of NMAR data is generated in which the missing-data mechanism is dependent on the true value of the outcome (i.e., the mechanism is dependent on θ_i, the value of the outcome before sampling error is added to the model, but after random-effects variation is added). This type of data (referred to as outcome-NMAR data, or o-NMAR data) is perhaps of the most concern to the skeptical researcher, who may worry that unmeasured variables both affect the outcome and relate to whether or not the effect size is reported. Consider, for instance, a research domain in which private schools are investigated disproportionally more often than public schools; in this instance, in comparison to the population of schools, data on public schools is missing more often than data on private schools. Further suppose that the public/private dimension has a strong

46

relationship to the outcome being investigated. If the researcher has no data on whether studies were conducted in public or private schools, her data will be NMAR in nature, dependent on the value of the outcome, because the outcome is dependent on the value of an unmeasured variable whose missingness is related to its value. In such circumstances one is assuming that the effect of unmeasured variables is subsumed in the value of the random-effects error u_i, and that the value of u_i is correlated to whether or not a study is observed (i.e., u_i is correlated to whether a study investigated a public or private school). Because the sampling error is unrelated to whether the school is public or private, the selection is based on the *population value* of the outcome, not the observed value of the outcome, which would include sampling error.

Another reason to investigate this type of missing-data mechanism is that in Little (1992), censoring based on the *observed value* of the outcome (which should lead to MAR data, as the outcome was always observed), showed poorer estimation of regression parameters than for an NMAR example. Further investigation is in order.

2. Generation of Data

Two series of simulations were conducted. The primary series crossed the hyperparameters listed above and was designed to investigate the bias and MSE of the complete-case, available-case, and maximum-likelihood estimators.

The second series was designed to test how well standard errors for the $\hat{\beta}$s could be estimated using the conditional bootstrapping method CBOOT described by Su (1988) when applied to maximum-likelihood estimation in this context.

Bias and MSE Simulations

For each of the 480 combinations of hyperparameters, 1000 simulated meta-analyses were generated and analyzed using the three methods (CC, AC, and ML estimation) described in Chapter 3. The values for the predictors and outcome were generated in the following fashion.

1.) The population βs are calculated based on which one of the two correlation matrices were chosen.

2.) Three columns of data, representing X_1, X_2, and X_3, are generated based on the correlation matrix. These three columns of data have the correlation matrix chosen as their population correlation matrix, though of course the actual sample correlation matrices among generated Xs will vary. Each column has a population mean of zero and a population variance of one.

3.) Variation stemming from variation in the predictors is added to the outcome.

$$\theta_i^* = \theta + (V_{mod})\beta_1 X_1 + (V_{mod})\beta_2 X_2 + (V_{mod})\beta_3 X_3 \, ,$$

where θ represents the selected population mean effect size. A mean effect size of .8 was used in all simulations.

4.) Variation stemming from random-effects error is added to the outcome.

$$\theta_i = \theta_i^* + u_i \, ,$$

where u_i represents the random-effects error for study i (var(u_i) = τ).

5.) Two sets of normally distributed raw data (y_is) with population mean effect size θ_i and sample size $n_i/2$ are generated. An unbiased standardized mean difference, T_i, is calculated from these data using the formula from Hedges & Olkin (1985), p. 81:

$$T_i = (1 - \frac{3}{4n_i - 9}) \frac{\overline{Y}_{1i} - \overline{Y}_{2i}}{s_i},$$

where s_i is the square root of the pooled sample variance.

6.) A missing data pattern is generated based on the methods described above for MCAR, MAR, p-NMAR, and o-NMAR data. This does not change the values of any of the data, but leads to some observations of some predictors to be considered "missing". The algorithm employed generates r.n.d.'s that (for all but the MCAR data) are correlated with the values of one of the predictors or the population value of the outcome. The algorithm then sorts the studies' r.n.d.'s and assigns each study a missing data pattern based on its value of the r.n.d. For instance, in the first missing data pattern, the studies with the 25% highest values were considered to be "complete", the studies with values in the middle 50% were considered to be "missing data on either the 2nd or 3rd predictor", and the studies with the 25% lowest values were considered to be "missing data on *both* the 2nd and 3rd predictor". As noted above and in Appendix 1, the algorithm is controlled so that studies within each missing data pattern have approximately the same proportion of small and large study sample sizes. Thus, *across* missing data patterns (i.e., complete-data, missing data on either the 2nd or 3rd predictor, or missing data on both 2nd and 3rd predictors), there is no relationship between average study sample size and the study population effect size. After each study was assigned a missing data pattern, the order of the studies within each pattern was randomized, in order to assure that *within* each missing data pattern there was no relationship between the average study sample size and the

study population effect size.

After the simulations were completed, the bias and mean-squared error (MSE) were calculated for the 1000 CC, AC, and ML estimators of β_0, β_1, β_2, β_3, and τ. These statistics were also calculated for the CC, AC, and ML estimators of the population mean. For the purpose of convenience in notation call a parameter of interest Θ and a given estimation technique's estimators of that parameter $\hat{\Theta}$.

The formula for bias in $\hat{\Theta}$ is Mean($\hat{\Theta}$) - Θ.

The formula for the MSE of $\hat{\Theta}$ is bias^2 + Var($\hat{\Theta}$).

The program to generate these data was written in SAS/IML Version 6.12 (SAS Institute, Inc.).

Standard Error Simulations

The purpose of these simulations was to test the accuracy of bootstrapped standard errors. The primary question was how many times (B) to run the bootstrap algorithm per meta-analysis. Efron and Tibshirani (1991) state that the benefit of increasing the value of B to over 200 is generally negligible, and that "values of B as small as 25 often give satisfactory results" (p. 391). Su (1988) used a value of B of only 2 in his simulated multiple regressions, determining that this was the optimal value after controlling for the total number of data sets and bootstrapped samples (see his Appendix B). Unfortunately, his method did not result in accurate bootstrap confidence intervals in each simulated regression, and he was forced to use additional calculations to calculate the coverage probabilities (see his Appendix C).

50

Given the considerable amount of computing time necessary to run the maximum-likelihood estimation procedure for these simulated meta-analyses, even a B of 25 is demanding for a large amount of simulation runs. However, I wanted to analyze the "real" coverage probabilities found using a higher B than the $B = 2$ method employed by Su (1988). Thus, a run was made of $B = 25$ for 100 simulated meta-analyses for each of 48 combinations of hyperparameters in which the data was MCAR and the correlation among the predictors was low. The incidence of missing data, the number of studies, the average study sample size, the size of the effect the predictors had on the outcome, and the size of the random-effects variance were all varied.

3. Criteria for the Investigation of Estimators

The estimators of most importance are $\hat{\beta}$ and $\hat{\tau}$. Though estimators of other parameters exist (e.g., of Σ_X and μ_X), estimation of these parameters is generally considered important only so far as it leads to good estimators of β and τ.

The accuracy of these estimators is first investigated by examining the estimators' biases. This is done in three ways: by examining the number of times the true value of the parameter fell outside the 99% confidence interval generated for its estimate (the 99% confidence interval was used given the large number of tests conducted and the high power of the tests), by looking at the actual size of the biases relative to the effect-size metric, and by calculating the ratios of the empirical MSEs to their empirical variances. The last method was employed by Su (1988) in his investigation of the bias of his MLE estimators, and allows one to judge the effect that the bias has relative to the error caused

51

by sampling error in the estimates. This ratio is expected to be 1.00 if no bias exists; if the ratio had a value of 1.10, this would imply that 1/1.10, or 91% of the variation in estimation stemmed from bias in the estimator.

After investigating the bias of each method's estimators of β and τ, the performance of ML and AC estimators are compared to the performance of the CC estimators by calculating the ratios of their estimators' respective MSEs. Thus, to compare the ML estimation of β_1 to the CC estimation of β_1, MSE_{CC}/MSE_{ML} is calculated for that estimator. This ratio is similar to the ratio calculated to measure "relative efficiency" (Mendenhall et al. 1986), but is not exactly said ratio because the ML estimators are not necessarily unbiased. However, in the presence of only a small bias, this ratio will be something very similar to relative efficiency. Even when there is bias, the ratio still fulfills a similar role.

These ratios were investigated by Su (1988) by comparing mean values through tables suggested by ANOVAs. A similar approach is used here to gauge the relative effects of simulation hyperparameters (e.g., the importance of variation in average study sample size versus the importance in variation in the population value of τ). However, ANOVA significance values are not reported, as they cannot be directly interpreted due to the two-level nature of the data. For instance, when the null hypothesis that the MSE of an EM estimator is the same as the MSE of a CC estimator is true for a given condition, the MSE_{CC}/MSE_{ML} ratio should be distributed as an F-statistic with 999 degrees of freedom in both the numerator and the denominator. The sampling error for such ratios is very low: the expected value is 999/997 = 1.002, and the standard error is approximately .062. Tests

52

of main effects, which as seen below may be based on as many as 196,000 observations, will have true standard errors of much less. This was confirmed by creating multiple random samples that were 25% the size of the largest dataset and comparing the average ratios across main effects. These averages changed by no more than .03 between these smaller subsamples.

ANOVAs do not take the standard errors of these ratios into account. Also, the standard error of the ratios will differ depending on what the true ratio actually is; the standard error for an $F_{999,999}$ will be .062 only when the null is true. It was decided to use ANOVAs only to investigate the relative size of effects, as mentioned above, and to pay close attention only to hyperparameters that had effects of at least .20.

The performance of the bootstrapped standard errors for the slopes was investigated in the two ways employed by Su (1988) in his investigation of bootstrapped standard errors. First, for each estimator, empirical non-coverage rates (using 95% confidence intervals) were calculated for the 100 simulated meta-analyses for each of the 96 combinations of hyperparameters mentioned above. In the second procedure, for each estimator, the average empirical variance over the 100 simulated meta-analyses was divided by the empirical MSE of the matching estimator. When this ratio equals one, the average estimated variance of the estimated slope is equal to the MSE of the estimated slopes, implying precise estimators of standard errors and an overall rejection rate which should be equal to that which is theoretically expected. For reasons discussed in the next chapter, a ratio was calculated that used the median empirical variance as well.

CHAPTER V

SIMULATION STUDY RESULTS

This investigation is split into six parts. The first part investigates the conditions for which the assumptions of the maximum-likelihood method is met (the data are MCAR or MAR because of missing data on the 1st predictor) when estimating β_0, β_1, β_2, β_3 and τ. In the second the assumptions are not met (the data are p-NMAR due to data being missing on the 2nd predictor), and in the third the assumptions are not completely met (the data are o-NMAR due to the data being missing on the true value of the always-observed outcome). The fourth part investigates all three of these missing data types regarding estimation of the population mean. The fifth part briefly considers the estimation of β_0, β_1, β_2, β_3 and τ when there is a dichotomous predictor with missing data. The final section investigates the performance of the bootstrapped standard errors.

1. Results: MCAR/MAR Data

For 192 combinations of hyperparameters the data were MCAR or MAR. These data are referred to jointly as MCAR/MAR data.

Bias in ML Estimation of β

It was expected that there would be no statistically significant bias in the MLE estimators, given the normally-distributed predictors and the almost normally distributed outcome. This was not the case. In the 192 conditions, there are 2 instances of bias for the

54

intercept, 10 instances for β_1, 25 instances for β_2, and 34 instances for β_3 ($p<.01$). There were 58 conditions for which there was a bias for at least one of the slopes. Of the seven hyperparameters varied in the simulation study, four are moderately related to whether a significant bias was found in the slopes: V_{mod} (for 39 of the conditions, $V_{mod} = .03$), average n_i (for 37 of the conditions, the average $n_i = 200$), missing data incidence (for 38 of the conditions, there was 50% missing data on both the 2nd and 3rd predictors, as opposed to 25%), and correlation matrix (for 38 of the conditions, the correlation matrix had low correlations among the predictors). There were no or very weak relationships between whether a condition showed a significant bias and that condition's values of k, τ, or whether the data was MCAR or MAR.

Table 5.1 contains the average, minimum, and maximum empirical MSE/variance ratios. A histogram of MSE/variance ratios for β_3 is in Figure 5.1. The third predictor was used to generate the figure as it is the 3rd slope for which there is the most evidence of bias. Both the table and the figure demonstrate that the biases are always very small relative to the amount of sampling error in the estimators. The largest ratio of 1.025 indicates that even the largest bias was only 2.5% the size of the normal sampling variation in the estimator.

Table 5.1

Maximum-Likelihood MSE/Variance Ratios for β (MCAR/MAR Data)			
	β_0	β_1	β_2	β_3
Mean	1.001	1.002	1.003	1.004
Median	1.001	1.001	1.001	1.002
Minimum	1.000	1.000	1.000	1.000
Maximum	1.011	1.021	1.025	1.024

Figure 5.1
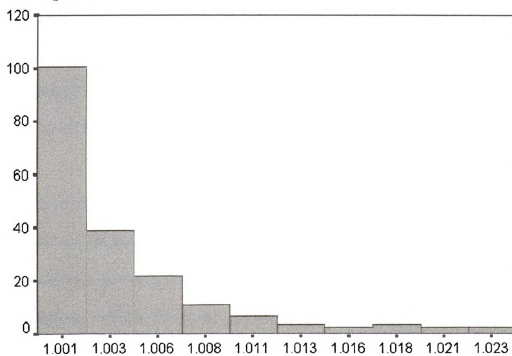
Ratio of MSE to Variance for β_3 (MCAR Data)

The presence of a substantively small, yet statistically significant bias is curious given that all assumptions of the maximum-likelihood model are met in these data except for the slightly non-normal outcome. Little and Rubin (1987) show that maximum-likelihood estimation should lead to consistent estimates, and Little (2000) said that while bias might exist in some more complex applications of maximum-likelihood theory, such as with censored data or in nonlinear regression (e.g., Cook, Tsai, & Wei, 1986), he was not familiar with instances of bias in the standard regression model with normal slopes and normal outcomes. He said that small biases might exist for small sample sizes, however, though they should go to zero as sample size increases. This view is reinforced in Cordeiro and McCullagh (1991), in which they state "It is well known that MLEs may be biased when the sample size or the total Fisher information is small. The bias is usually ignored in practice, the justification being that it is negligible compared with the standard errors." (p. 629)

The work of Su (1988) supports the idea that in some cases, there might be some bias in maximum-likelihood estimates in the standard general linear model even when the outcome and predictors are normally distributed. Su found MSE ratios of 1.05 and 1.02 for his sample size of 40, and 1.00 and 1.01 for his sample sizes of 160, and his model did not concern itself with a known fixed covariance term of different sizes across studies or treat the random-effects variation the same as the ML estimation in Chapter 3. These numbers compare favorably with what is found in the present study.

Follow-up simulations were conducted in which $k=400$ and a small number for which $k=1000$. In approximately 2/3 of the cases, the bias shrunk to become statistically

57

insignificant. In the other 1/3, however, the bias shrunk no faster than the standard error of the estimates. In other words, while the bias shrunk as might be expected if the ML estimates were consistent, there were still statistically significant biases for some sets of conditions for higher values of k. In some of this simulation work a normal outcome was simulated, and equally-sized studies; neither of these changes affected the size of the bias. Because the MSE ratio was still small (1.01 to 1.02) and the biases for the higher values of k so substantively uninteresting (on the order of one-hundredth of an effect size) the matter of the bias was left to later investigation.

Bias in ML Estimation of τ

Some bias was expected in the estimation of τ given that the method used was full maximum-likelihood and not restricted maximum-likelihood (Bryk & Raudenbush, 1992, p. 223). The size of the bias was expected to be lower for $k = 100$ than for $k = 40$. An ANOVA showed that all seven simulation hyperparameters and most 2-way interactions between those hyperparameters were significant; however, the only substantive effects came from missing data pattern, number of studies, and the value of τ. Table 5.2 shows the ratios and biases for the ML estimations across the different values of these simulation hyperparameters. The average ratio across all conditions was 1.327.

Table 5.2

Maximum-Likelihood MSE/Variance Ratios for τ (MCAR/MAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	k = 40	τ = 0	1.179	.0012
		τ = .005	1.276	-.0020
		τ = .02	1.398	-.0075
	k = 100	τ = 0	1.380	.0012
		τ = .005	1.117	-.0011
		τ = .02	1.111	-.0031
75% Incidence of Missing Data	k = 40	τ = 0	1.148	.0011
		τ = .005	1.608	-.0028
		τ = .02	1.809	-.0110
	k = 100	τ = 0	1.342	.0016
		τ = .005	1.273	-.0017
		τ = .02	1.284	-.0054

While the ratios are sometimes quite large (above 2.00 for some combinations of hyperparameters), the size of the bias is usually trivial. The only exception is when τ = .02 and the number of studies is small. In these instances biases average -.011, roughly half the size of the population variance to be estimated. As expected, this bias drops when study size increases or more data are observed.

Bias in AC Estimation of β

As noted in the review of literature, some biases in the AC estimates of β were not

unexpected, though it was unknown how large they might be. Table 5.3 shows how often there were statistically significant ($p<.01$) biases for the estimates of β. MAR data led to more biases than MCAR data, especially for β_0 and β_1.

Table 5.3

Available-Case Frequencies of Bias in Estimates of β (MCAR/MAR Data)				
Missing-data Mechanism	β_0	β_1	β_2	β_3
MCAR	8 (8.3%)	21 (21.9%)	31 (32.2%)	39 (40.6%)
MAR	74 (77.1%)	51 (53.1%)	40 (41.7%)	58 (60.4%)

The amount of bias in the AC estimators differed greatly depending on the simulation patterns. Not surprisingly, bias was the worst when there was the least complete data (75% missing data and $k = 40$). The size of the bias varied widely for this condition and often was affected by outliers across the 1000 simulations.

While bias was widespread, the size of the bias in $\hat{\beta}$ relative to the variance in $\hat{\beta}$ was typically small. Tables 5.4 and 5.5 summarize the ratios of MSEs to the variances of the estimators. For both the MCAR and MAR data, the biases in $\hat{\beta}$ are on average substantively insignificant relative to the mean-squared error of estimation.

Table 5.4

Available-Case MSE/Variance Ratios for β (MCAR Data)						
		β_0	β_1	β_2	β_3
50% Incidence	k=40	1.000	1.002	1.002	1.005
of M. Data	k=100	1.001	1.001	1.002	1.002
75% Incidence	k=40	1.001	1.002	1.002	1.007
of M. Data	k=100	1.001	1.002	1.002	1.006

Table 5.5

Available-Case MSE/Variance Ratios for β (MAR Data)						
		β_0	β_1	β_2	β_3
50% Incidence	k=40	1.103	1.013	1.003	1.005
of M. Data	k=100	1.327	1.042	1.005	1.008
75% Incidence	k=40	1.012	1.002	1.003	1.007
of M. Data	k=100	1.040	1.004	1.004	1.007

The only exception is the estimation of β_0 and β_1 for the MAR data, especially when there is little missing data and there are a large number of studies. In one set of conditions the ratio was as high as 3.00. It should be noted that the reason the MAR ratios are so high in these cases is not because the biases are higher, but because the variance of the estimates is much lower, due to the large number of simulated studies or the low percentage of missing data. The ratios in the table are averages; however, for no set of

MCAR conditions does an estimator of any β have a ratio above 1.035. For no set of MAR conditions (where there is 75% missing data) does the ratio exceed 1.20.

Bias in AC Estimation of τ

Given the theoretical possibility of bias in $\hat{\beta}$ it was expected that there might be some bias in the estimation of τ, especially given the fact that all negative estimates of τ were set to zero (as they would be in a real study). Table 5.6 summarizes the sizes of the average bias in $\hat{\tau}$ for the hyperparameters that were most strongly related to the size of bias. The estimation is the worst for $\tau = 0$; the MSE/variance ratios are high and the actual size of the bias is large, especially for average $n_i = 80$. The bias for these conditions is especially large; an average bias of .0078 for population $\tau = 0$ implies a standard deviation in the outcome of .088, which could be substantively misleading. There are similarly substantively large biases for $\tau = .005$ when average $n_i = 80$. The bias is negligible relative to the size of τ for $\tau = .02$ across all conditions.

Table 5.6

Available-Case MSE/Variance Ratios and Biases for τ (MCAR/MAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.387	.0078
		$\tau = .005$	1.147	.0056
		$\tau = .02$	1.016	.0022
	Average $n_i = 400$	$\tau = 0$	1.325	.0026
		$\tau = .005$	1.011	.0007
		$\tau = .02$	1.011	-.0010
75% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.275	.0073
		$\tau = .005$	1.080	.0047
		$\tau = .02$	1.001	-.0001
	Average $n_i = 400$	$\tau = 0$	1.246	.0026
		$\tau = .005$	1.003	.0001
		$\tau = .02$	1.047	-.0027

Bias in CC Estimation of β and τ

As expected by estimation theory, there was no bias in the complete-case

estimators of β for the MCAR or MAR data. Because all negative estimates of τ were set

to zero before means and MSE/variance ratios were calculated, some positive bias was

expected; however, biases were generally less than .003, and no ratio exceeded 1.025.

MSE$_{CC}$ to MSE$_{MLE}$ Ratios

Due to the lack of bias among over 70% of the simulation conditions, and the relatively small bias present for some predictors for the remaining 30%, it is fair to roughly characterize the MSE$_{CC}$/MSE$_{MLE}$ ratios below as measuring efficiencies of the ML estimators relative to the CC estimators. One ratio was calculated for the intercept, one ratio was calculated for each slope, and one ratio was calculated for τ.

Table 5.7 summarizes the MSE ratios across all simulation conditions. On average, the maximum-likelihood analysis provides large gains in efficiency; estimation of the intercept and slope for the completely-observed variable is almost twice as efficient, and estimation of τ over three-and-a-half times as efficient. However, while for each parameter to be estimated the maximum-likelihood method is more efficient than the complete-case method across all conditions, the relative efficiency varies considerably depending on the combinations of hyperparameters.

Table 5.7

MSE$_{CC}$/MSE$_{MLE}$ Ratios (MCAR/MAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.927	1.863	1.299	1.332	3.584
Minimum	1.162	1.183	1.023	1.061	1.090
Maximum	4.060	3.366	1.986	1.995	14.510

All of the lowest relative efficiencies came from studies with an average study sample size of 400. The lowest relative efficiencies for β_0 and β_1 came from the set of

conditions with $\tau=0$, $k=100$, $V_{mod}=.03$, low correlation among predictors, and 50%

missing MCAR data. The lowest relative efficiency for τ came from a similar set of

hyperparameters, except that $\tau = .005$. The lowest relative efficiencies for estimation of

β_2 and β_3 came from somewhat different looking conditions: they had in common $\tau =$

.02, $k=40$, high correlations among predictors, and a 75% incidence of missing data per

predictor. For β_2, however, $V_{mod} = .03$ and the data was MAR; for β_3, $V_{mod} = .006$ and

the data was MCAR. For none of these sets of conditions was there any bias on any of the

slopes except for a small bias in β_2 for the τ minimum relative efficiency. This supports

the conclusion made above that the slight bias in the estimates of β were substantively

insignificant.

Table 5.8 below summarizes the MSE ratios for the main effects. The effect of

different sizes of τ is substantively trivial for relative efficiency in estimation of β_0 and β_1,

moderately affects relative efficiency for β_2 and β_3, and has a large effect on relative

efficiency in the estimation of τ. Overall, having more studies in a meta-analysis leads to

slightly lower efficiencies for the estimation of the βs, but leads to clearly better complete-

case estimates of τ, and thus a smaller gain in efficiency. The same is true regarding

having a larger average study sample size, but the effect on the relative efficiencies of the

βs is slightly larger and the effect on the efficiencies for τ is slightly smaller.

The correlation among the predictors also has a small effect. Higher

intercorrelations among predictors lead to smaller gains in relative efficiency for the βs but

a larger gain in relative efficiency for estimation of τ. Incidence of missing data is the most

important hyperparameter with regard to its effect on relative efficiency.

Table 5.8

MSE$_{CC}$/MSE$_{MLE}$ Ratios, Main Effects (MCAR/MAR Data)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.969	1.894	1.431	1.456	4.700
	.005	1.921	1.836	1.286	1.322	2.572
	.02	1.892	1.781	1.181	1.220	1.694
k	40	1.974	1.909	1.329	1.353	3.657
	100	1.880	1.765	1.269	1.311	2.321
Avg. n_i	80	2.040	1.943	1.348	1.362	3.429
	400	1.815	1.731	1.250	1.302	2.549
Predictor inter-corrs.	Low	2.052	1.921	1.333	1.367	2.825
	High	1.802	1.752	1.265	1.298	3.152
V_{mod}	.006	2.036	1.931	1.320	1.342	3.297
	.03	1.818	1.743	1.278	1.323	2.681
Incidence of M. Data	50%	1.596	1.551	1.188	1.209	2.248
	75%	2.258	2.123	1.411	1.456	3.729
M. Data Mechanism	MCAR	1.677	1.691	1.303	1.330	3.011
	MAR	2.178	1.983	1.295	1.334	2.967

It has the largest main effect for each of the βs, and the second largest main effect for the estimation of τ; predictably, on average, the more missing data there is, the better the maximum-likelihood method performs relative to the complete-case method. Finally, the type of missing-data mechanism (MCAR vs. MAR) has (large) substantive main effects on

the efficiencies for β_0 and β_1, but little effect on β_2, β_3, or τ. Somewhat surprisingly, even though the estimation of the βs is unbiased for complete-case analysis in the presence of MAR data, the estimation of the parameters is clearly worse than it is for MCAR data. Note that the efficiencies for β_2 and β_3 are quite similar. This pattern was maintained through most of the simulation study. The values of β_2 and β_3 were not identical, but because both variables X_2 and X_3 were missing the same amount of the time, statistics regarding these two parameters behaved in similar ways, regarding interaction effects, relative efficiencies, and so on.

Although most that can be learned about the relative efficiencies can be gained from Table 5.8, there were two noticeable 2nd-order interaction effects for the MSEs for β_0 and β_1. These are shown in Tables 5.9 and 5.10. These tables show that the ML estimation of β_0 and β_1 is especially efficient relative to CC estimation when there are low predictor intercorrelations and a large amount of missing data, or low predictor intercorrelations and MAR data. There were no interesting (i.e., substantively large) interactions for the MSEs for β_2 and β_3.

There are several large interaction effects for the MSE ratios for the estimation of τ, the largest of which involve the size of τ and another hyperparameter. These interactions are reflected in Tables 5.11 - 5.14.

Table 5.11 demonstrates that the interaction between the size of τ and the incidence of missing data is such that when there is 75% missing data, the ML method is especially efficient for low values of τ, especially $\tau = 0$.

Table 5.12 shows that for a small average study sample size, the ML method is

especially efficient for $\tau = 0$ and $\tau = .005$. For a larger average study sample size, the ML method is only especially efficient (i.e., three to four times as efficient) for $\tau = 0$.

In Table 5.13 the key facet of the interaction is that the ML method is especially efficient (almost six times as efficient) for a small number of studies and a low value of τ.

Table 5.14 demonstrates that the ML method is especially efficient when there is the least amount of complete data, i.e., when there is a 75% incidence of missing data and only 40 studies.

Table 5.9

MSE$_{CC}$/MSE$_{MLE}$ Ratios for β_0, MCAR/MAR Data (Predictor Intercorrelations x Missing Data Incidence)		Ratio for β_0	Ratio for β_1
Low Predictor Intercorrelations	50% Incidence of Missing Data	1.568	1.559
	75% Incidence of Missing Data	2.537	2.284
High Predictor Intercorrelations	50% Incidence of Missing Data	1.625	1.542
	75% Incidence of Missing Data	1.978	1.963

Table 5.10

MSE$_{CC}$/MSE$_{MLE}$ Ratios for β_0, MCAR/MAR Data (Predictor Intercorrelations x Missing-Data Mechanism)		Ratio for β_0	Ratio for β_1
Low Predictor Intercorrelations	MCAR	1.647	1.697
	MAR	2.457	2.145
High Predictor Intercorrelations	MCAR	1.706	1.685
	MAR	1.899	1.820

Table 5.11

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR/MAR Data (Size of τ x Incidence of Missing Data)		
		Ratio for τ
50% Missing Data	$\tau = 0$	3.375
	$\tau = .005$	1.922
	$\tau = .02$	1.447
75% Missing Data	$\tau = 0$	6.026
	$\tau = .005$	3.222
	$\tau = .02$	1.940

Table 5.12

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR/MAR Data (Size of τ x Average Study Sample Size)		
		Ratio for τ
Average $n_i = 80$	$\tau = 0$	4.857
	$\tau = .005$	3.518
	$\tau = .02$	1.911
Average $n_i = 400$	$\tau = 0$	4.544
	$\tau = .005$	1.625
	$\tau = .02$	1.477

Table 5.13

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR/MAR Data		
(Size of τ x Number of Studies)		
		Ratio for τ
$k = 40$	$\tau = 0$	5.880
	$\tau = .005$	3.203
	$\tau = .02$	1.887
$k = 100$	$\tau = 0$	3.521
	$\tau = .005$	1.941
	$\tau = .02$	1.500

Table 5.14

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR/MAR Data		
(Incidence of Missing Data x Number of Studies)		
		Ratio for τ
50% Missing Data	$k = 40$	2.602
	$k = 100$	1.894
75% Missing Data	$k = 40$	4.731
	$k = 100$	2.747

Unlike for CC estimation and ML estimation, there is no theory that suggests that AC estimation is appropriate when the data are MAR. However, the results of ANOVAs suggested that the performance of the AC estimation with MAR data was very similar to its performance with MCAR data; thus, the results for the two types of data are examined simultaneously. The few differences that existed are noted below.

As will be seen below, except for the estimation of τ, and except for the estimation of the βs when $k = 40$ and there was 75% incidence of missing data, the biases among the AC estimators are small enough that it is fair to roughly characterize the MSE$_{CC}$/MSE$_{AC}$ ratios below as measuring the efficiencies of the AC estimators relative to the CC estimators.

Table 5.15 summarizes the MSE ratios across all simulation conditions.

Table 5.15

MSE$_{CC}$/MSE$_{AC}$ Ratios (MCAR/MAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.424	1.286	.862	.866	1.155
Median	1.343	1.321	.942	.939	1.072
Minimum	.010	.014	.007	.010	.081
Maximum	3.539	2.751	1.402	1.433	2.664

On average, available-case methods provide much better estimates than complete-

case methods for β_0 and β_1, slightly better estimates for τ, and slightly worse estimates for β_2 and β_3. However, the results have high variability across the values of the hyperparameters. Table 5.16 summarizes the MSE ratios by main effect. The largest effects are for the size of correlation among predictors and the number of studies per meta-analysis. The only substantively interesting interaction effects for the βs involve incidence of missing data, the correlation among the predictors, and the number of studies in the simulated meta-analysis. Table 5.17 shows the average sizes of these interactions across the combinations of the different values of these hyperparameters. The interaction between the correlation among predictors and missing data incidence is especially strong for β_0 and β_1, as is the interaction between incidence of missing data and the number of studies for β_2 and β_3. Available-case estimation of β_0 and β_1 is especially poor for $k = 40$ when there are high intercorrelations between predictors, and AC estimation of β_2 and β_3 is especially poor for $k = 40$ when the incidence of missing data is high.

Table 5.16

MSE$_{CC}$/MSE$_{AC}$ Ratios, Main Effects (MCAR/MAR Data)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.226	1.137	.772	.771	1.053
	.005	1.469	1.325	.888	.891	1.142
	.02	1.577	1.395	.928	.937	1.270
k	40	1.325	1.114	.736	.737	1.254
	100	1.523	1.458	.989	.996	1.056
Avg. n_i	80	1.521	1.371	.907	.907	1.349
	400	1.327	1.200	.818	.826	.960
Predictor inter-corrs.	Low	1.654	1.479	.968	.981	1.179
	High	1.194	1.093	.757	.752	1.131
V_{mod}	.006	1.512	1.348	.873	.872	1.219
	.03	1.337	1.224	.852	.861	1.090
Incidence of M. Data	50%	1.272	1.285	.939	.938	.932
	75%	1.576	1.287	.786	.795	1.377
M. Data Mechanism	MCAR	1.295	1.267	.982	.988	1.063
	MAR	1.626	1.563	.974	.977	1.038

Table 5.17

MSE$_{CC}$/MSE$_{AC}$ Ratios, MCAR/MAR Data (Incidence of Missing Data x *k* x Predictor Intercorrelations)			β_0	β_1	β_2	β_3
50% Incidence of M. Data	*k* = 40	Low Corrs.	1.382	1.400	1.025	1.032
		High Corrs.	1.291	1.261	.888	.878
	k = 100	Low Corrs.	1.270	1.279	.963	.970
		High Corrs.	1.147	1.204	.878	.873
75% Incidence of M. Data	*k* = 40	Low Corrs.	1.837	1.345	.722	.738
		High Corrs.	.792	.454	.697	.684
	k = 100	Low Corrs.	2.127	1.897	1.160	1.183
		High Corrs.	1.548	1.451	.956	.959

Table 5.18

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR/MAR Data (Size of τ x Number of Studies)		τ Ratio
Average n_i = 80	$\tau = 0$	1.201
	$\tau = .005$	1.212
	$\tau = .02$	1.254
Average n_i = 400	$\tau = 0$	.660
	$\tau = .005$	.833
	$\tau = .02$	1.141

The patterns of ratios for the βs are roughly what would be expected given what we know about pairwise estimates' sensitivity to extreme correlation matrices. The pairwise estimates have the highest relative efficiency when the correlations among the variables are the lowest, and especially when the correlations between the predictors and the outcome are lowest – specifically, when τ is large, V_{mod} is low, or the average sample size is low. However, on average the ratios are much closer to 1.00 than they were for the maximum-likelihood analyses. Just as was the case with the ML/CC relative efficiencies, available-case analyses show greater gains in efficiency for MAR data than for MCAR data for estimation of β_0 and β_1.

There is one moderate interaction between average n_i and τ; it is summarized in Table 5.18. The interaction represented in Table 5.18 shows that the effect of study sample size is relatively constant across values of τ for average $n_i = 80$, but graduated for average $n_i = 400$.

Because, in general, available-case estimation does not perform better than complete-case estimation for β_2 and β_3, it might be concluded that pairwise estimation is not recommended for use in a mixed-model meta-analysis with MCAR or MAR data. However, there are large gains in efficiency for β_0 and β_1 when there is a large amount of random-effects variation in the studies, and there are either a large number of studies or a small percentage of missing data. Under these conditions, AC estimation compares favorably with even the maximum-likelihood method, as shown in Table 5.19. While maximum-likelihood estimation of τ is on average 26% more efficient for even these conditions, the estimation of the βs is on average only 10 to 13 percent more efficient, and

on a few rare occasions (five conditions out of the 192) is marginally less efficient than AC estimation. These cases arose when V_{mod} = .006 and the correlations among the predictors were at their lowest (.1).

Table 5.19

MSE_{AC}/MSE_{MLE} Ratios for MCAR/MAR Data, τ = .02 (Excluding k=40/75% Missing Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.121	1.108	1.100	1.134	1.278
Median	1.100	1.090	1.096	1.136	1.288
Minimum	1.013	.991	.927	.921	1.116
Maximum	1.629	2.086	1.598	1.746	1.470

Thus, the ML estimation of the βs does the most poorly relative to AC analysis when there are almost zero correlations between the predictors and outcome, and very low correlations among predictors. It should be noted that in these cases, complete-case analysis is rarely appropriate; both the ML and AC estimation procedures are on average two to three times as efficient in estimating β_0 and β_1, and 30% to 40% as efficient in estimating β_2 and β_3.

2. Results: p-NMAR Data

There were a total of 192 combinations of hyperparameters for which the data were NMAR and the missing-data mechanism was dependent on the values of the predictors. 96 of these combinations had data that were strongly NMAR, (sp-NMAR), while the other 96 had combinations of hyperparameters that were moderately NMAR (mp-NMAR).

Bias in ML Estimation of β

It was expected that there would be statistically significant biases in the MLE estimators of the slopes given that ML theory assumes that the missing-data mechanism is MCAR or MAR. This proved to be the case, as shown in Table 5.20. Table 5.21 shows the average, minimum, and maximum MSE/variance ratios across all data.

The frequency of bias is low for the slopes for the predictors, although the intercept is almost always biased. The primary difference between these results and the results for the MCAR/MAR data in the previous section is the frequency of bias in the intercept. This finding is borne out in histograms of the sizes of the biases. Figures 5.2 and 5.3 are of the biases for β_3. Except for some rare cases, the ratios are very close to 1.00. The highest ratios come from cases with $\tau = .005$, 75% incidence of missing data, low correlations among predictors, and an average study sample size of 400. The actual size of the significant biases for the slopes ranged from -.06 to .09. While the larger biases are not substantively uninteresting, these figures show they are generally overwhelmed by sampling error.

Table 5.20

Maximum-Likelihood Frequencies of Bias in Estimates of β (p-NMAR Data)				
Missing-data Mechanism	β_0	β_1	β_2	β_3
sp-NMAR	95(99%)	28(29.2%)	13(13.5%)	38(29.2%)
mp-NMAR	89(92.7%)	18(18.8%)	11(11.5%)	20(20.8%)

Table 5.21

Maximum-Likelihood MSE/Variance Ratios for β (p-NMAR Data)				
	β_0	β_1	β_2	β_3
Mean	1.137	1.005	1.003	1.007
Median	1.078	1.002	1.001	1.002
Minimum	1.002	1.000	1.000	1.000
Maximum	1.740	1.021	1.084	1.087

Figure 5.2
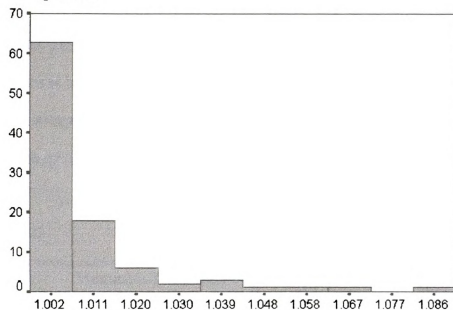
Ratio of MSE to Variance for β_3 (sp-NMAR Data)

Figure 5.3
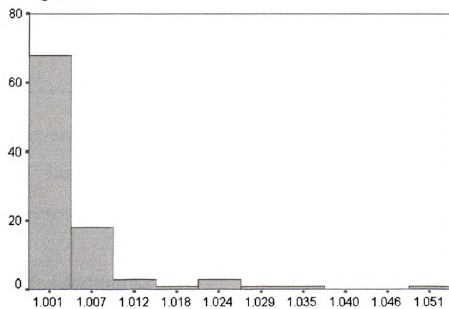
Ratio of MSE to Variance for β_3 (mp-NMAR Data)

The ratios for the biases for the intercept are on average much higher than for the slopes, but substantively they are small, as shown in Table 5.22. While the ratios are sometimes as high as 1.74, the maximum bias is only -.0297. Given the relative unimportance of the difference between an average effect size of .8 versus an average effect size of .77, the bias in the intercept often can be safely ignored.

Table 5.22

Maximum-Likelihood MSE/Variance Ratios and Biases for β_0 and β_1 (p-NMAR Data)				
			β_0 Ratio	Bias in $\hat{\beta}_0$
sp-NMAR Data		Mean	1.204	-.0139
		Median	1.117	-.0125
		Minimum	1.003	-.0034
		Maximum	1.740	-.0297
mp-NMAR Data		Mean	1.070	-.0072
		Median	1.052	-.0068
		Minimum	1.002	-.0018
		Maximum	1.246	-.0177

Bias in ML Estimation of τ

An ANOVA in which the outcome was the size of the bias in the estimation of τ was conducted that showed that all seven simulation hyperparameters and most 2-way interactions between those hyperparameters were significant; however, the only

substantive effects came from missing data pattern, number of studies, and the value of τ. Whether the data were sp-NMAR or mp-NMAR was unimportant. Table 5.23 shows the ratios and biases for the ML estimates. All biases were in a positive direction. The large negative biases for $\tau = .02$ found for the MCAR/MAR data are not present in the NMAR estimations. However, there are substantively large biases for $k = 40$ and 75% incidence of missing data for the lower values of τ.

Table 5.23

Maximum-Likelihood MSE/Variance Ratios for τ (p-NMAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	$k = 40$	$\tau = 0$	1.177	.0035
		$\tau = .005$	1.274	.0028
		$\tau = .02$	1.364	.0019
	$k = 100$	$\tau = 0$	1.386	.0022
		$\tau = .005$	1.114	.0017
		$\tau = .02$	1.104	.0015
75% Incidence of Missing Data	$k = 40$	$\tau = 0$	1.149	.0065
		$\tau = .005$	1.561	.0056
		$\tau = .02$	1.824	.0046
	$k = 100$	$\tau = 0$	1.332	.0031
		$\tau = .005$	1.231	.0025
		$\tau = .02$	1.242	.0016

Bias in AC Estimation of β

The extent to which biases were statistically significant in the sp-NMAR and mp-NMAR data is shown below in Table 5.24.

Table 5.24

Available-Case Frequencies of Bias in Estimates of β (p-NMAR Data)				
Missing-data Mechanism	β_0	β_1	β_2	β_3
sp-NMAR	96 (100%)	40 (41.7%)	71 (74.0%)	64 (66.7%)
mp-NMAR	93 (96.9%)	36 (37.5%)	40 (41.7%)	41 (42.7%)

While bias was widespread, the size of the bias relative to the variance of the estimators was similar to that found for the MAR data for all estimators but β_0. Tables 5.25 and 5.26 summarize the ratios of MSEs to the variances of the estimators. The biases for the βs are, on average, substantively insignificant relative to the mean-squared error for both the MCAR and MAR data for β_2 and β_3, especially for the mp-NMAR data. In contrast, the ratios for β_0 were quite large, indicating that much of the error in the estimation of β_0 came from biased estimation.

Table 5.25

Available-Case MSE/Variance Ratios for β (sp-NMAR Data)					
		β_0	β_1	β_2	β_3
50% Incidence	k=40	1.361	1.006	1.018	1.009
of M. Data	k=100	2.062	1.019	1.042	1.018
75% Incidence	k=40	1.228	1.001	1.008	1.010
of M. Data	k=100	2.592	1.021	1.038	1.018

Table 5.26

Available-Case MSE/Variance Ratios for β (mp-NMAR Data)					
		β_0	β_1	β_2	β_3
50% Incidence	k=40	1.097	1.003	1.004	1.004
of M. Data	k=100	1.247	1.002	1.004	1.003
% Incidence of	k=40	1.117	1.004	1.007	1.013
M. Data	k=100	1.515	1.003	1.006	1.005

Table 5.27 summarizes the sizes of the average ratios and bias for β_0 across the types and incidence of missing data, and values of V_{mod}. These variables were the most important in predicting the amount of bias. Biases are larger for $V_{mod} = .03$, the sp-NMAR data, and 75% missing data. There were no substantively large interactions. While the biases are larger than those for the ML estimator of β_0 (shown in Table 5.22), on average they are small relative to the effect size metric.

84

Table 5.27

Available-Case MSE/Variance Ratios and Biases for β_0 (p-NMAR Data)				
			Ratio for β_0	Bias in $\hat{\beta}_0$
sp-NMAR Data	50% Incidence of Missing Data	$V_{mod} = .006$	1.246	-.0150
		$V_{mod} = .03$	2.177	-.0350
	75% Incidence of Missing Data	$V_{mod} = .006$	1.362	-.0320
		$V_{mod} = .03$	2.458	-.0560
mp-NMAR Data	50% Incidence of Missing Data	$V_{mod} = .006$	1.069	-.0075
		$V_{mod} = .03$	1.274	-.0160
	75% Incidence of Missing Data	$V_{mod} = .006$	1.120	-.0130
		$V_{mod} = .03$	1.512	-.0290

Bias in AC Estimation of τ

Table 5.28 summarizes the sizes of the average bias in $\hat{\tau}$ for the hyperparameters that were most strongly related to the size of bias. The results are very similar to those for the MCAR/MAR data. Estimation is the worst for $\tau = 0$, and biases are substantively large for the conditions in which $\tau = 0$ or .005 and the average $n_i = 80$. Biases are substantively negligible for other conditions.

Table 5.28

Available-Case MSE/Variance Ratios and Biases for τ (p-NMAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.406	.0081
		$\tau = .005$	1.158	.0060
		$\tau = .02$	1.014	.0023
	Average $n_i = 400$	$\tau = 0$	1.343	.0025
		$\tau = .005$	1.017	.0009
		$\tau = .02$	1.007	-.0007
75% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.324	.0074
		$\tau = .005$	1.099	.0056
		$\tau = .02$	1.006	.0002
	Average $n_i = 400$	$\tau = 0$	1.277	.0025
		$\tau = .005$	1.007	.0004
		$\tau = .02$	1.004	-.0021
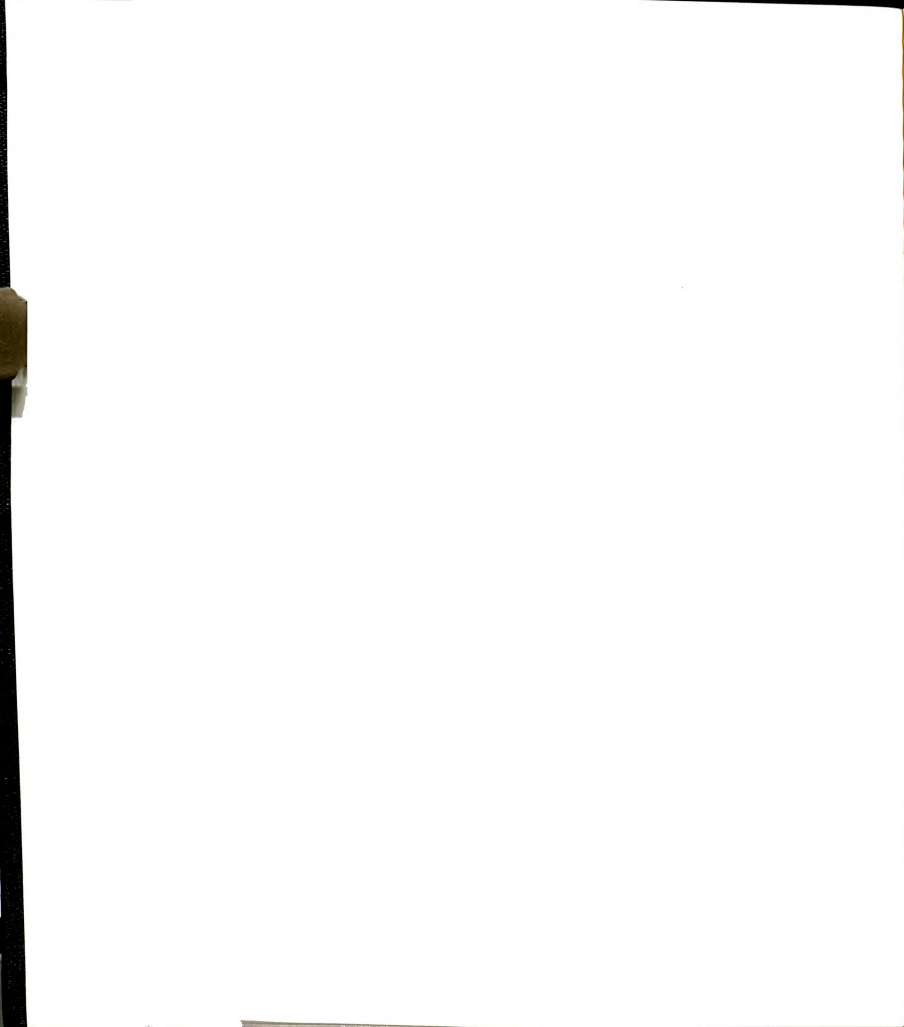
Bias in CC Estimation of β and τ

Out of the 192 conditions, there were only 7 instances of a significant bias for estimation of β_0, 8 instances of significant bias for β_1, 8 instances of significant bias for β_2, and 4 instances of significant bias for β_3. These biases were very small relative to the sampling error; the ratios of MSEs to variance were all less than 1.02.

Bias in the estimation of τ was more substantial. Table 5.29 summarizes the sizes of the bias over the simulation conditions to which it was most strongly related: τ, k, incidence of missing data, and average n_i. Biases in the estimation of low values of τ were very large, sometimes over .01, in cases where the average study sample size was small. There were substantial interaction effects between average study sample size and both incidence of missing data and number of studies. A high proportion of missing data or low value for k exacerbated the bias in the estimation of τ for the conditions in which average study sample size was 80.

Table 5.29

Complete-Case MSE/Variance Ratios and Biases for τ (p-NMAR Data)			
		Ratio for τ	Bias in $\hat{\tau}$
$\tau = 0$	Average $n_i = 80$	1.383	.0113
	Average $n_i = 400$	1.363	.0021
$\tau = .005$	Average $n_i = 80$	1.187	.0096
	Average $n_i = 400$	1.026	.0011
$\tau = .02$	Average $n_i = 80$	1.044	.0067
	Average $n_i = 400$	1.004	.0005
$k = 40$	Average $n_i = 80$	1.202	.0126
	Average $n_i = 400$	1.122	.0018
$k = 100$	Average $n_i = 80$	1.208	.0058
	Average $n_i = 400$	1.140	.0007
50% Incidence of Missing Data	Average $n_i = 80$	1.202	.0120
	Average $n_i = 400$	1.127	.0017
75% Incidence of Missing Data	Average $n_i = 80$	1.209	.0064
	Average $n_i = 400$	1.134	.0008

Table 5.30 summarizes the MSE ratios across all simulation conditions. As it did with the MCAR/MAR data, the maximum-likelihood analysis provides large gains in efficiency; estimation of the intercept and slope for the completely-observed variable is almost twice as efficient, and estimation of τ is almost three times as efficient.

Table 5.30

MSE$_{CC}$/MSE$_{MLE}$ Ratios (p-NMAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.871	1.677	1.417	1.306	2.861
Median	1.764	1.536	1.347	1.246	2.373
Minimum	1.107	1.212	1.061	1.038	1.044
Maximum	3.383	2.559	2.463	2.388	10.73

As with the MCAR/MAR data, the relative efficiency varied depending on the simulation conditions. Table 5.31 shows the main effects of each of the seven simulation hyperparameters. For the intercept and slopes, the only hyperparameters that were very important were the incidence of missing data and the population value of τ. The lower the value of τ, the higher the relative efficiency; similarly, the more missing data there were, the higher the relative efficiency was. The effects for the other five hyperparameters on the estimation of the βs was slight or non-existent. More hyperparameters affected the estimation of τ. The population value of τ, the average study sample size, the number of
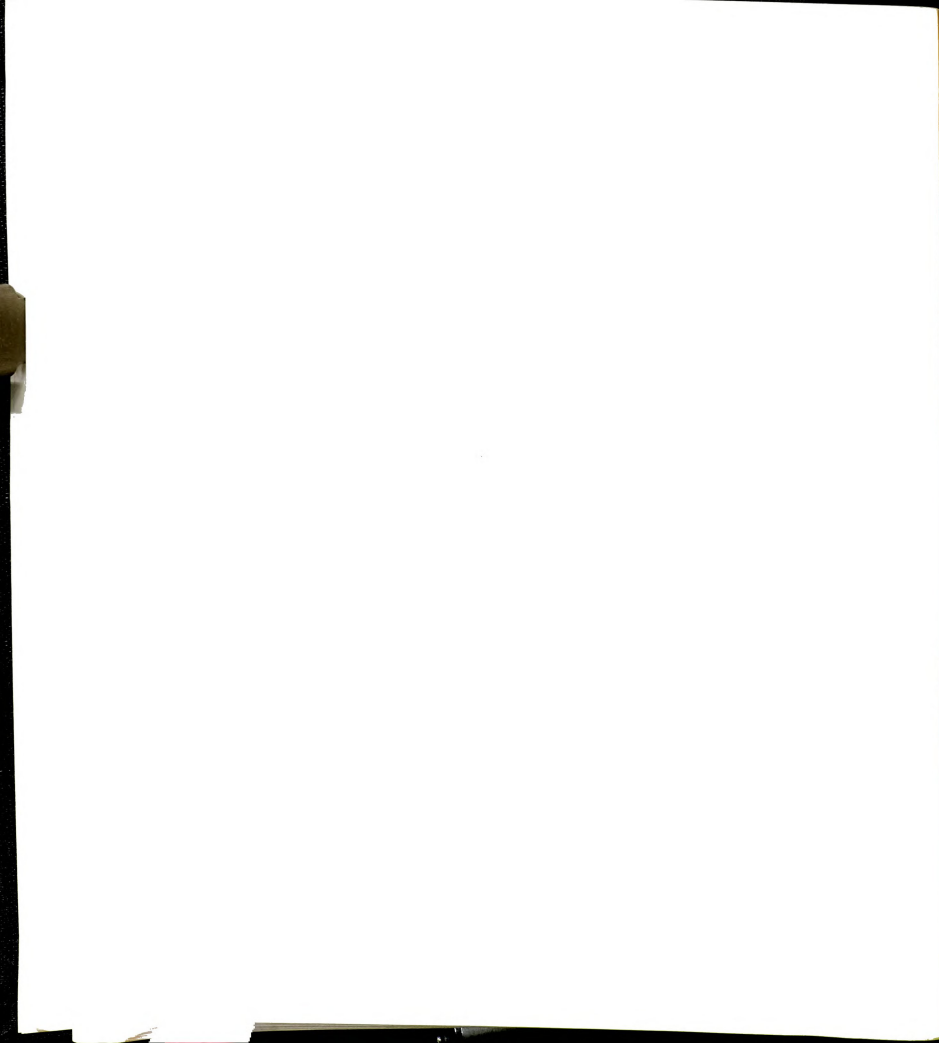
studies, the value of V_{mod}, and the incidence of missing data all had moderate to large effects on the relative efficiency. There was little to no effect for whether the missing-data mechanism was sp-NMAR or mp-NMAR on the efficiencies for any of the hyperparameters. The relative efficiencies for $\hat{\beta}_2$ and $\hat{\beta}_3$ were quite similar to each other, even though there was an important difference between the 2nd and 3rd predictors in this combination of simulation hyperparameters – the missing-data mechanism was related to a function of solely the 2nd predictor.

Table 5.31

MSE$_{CC}$/MSE$_{MLE}$ Ratios, Main Effects (p-NMAR Data)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.952	1.746	1.550	1.430	4.417
	.005	1.862	1.676	1.404	1.292	2.490
	.02	1.799	1.610	1.295	1.194	1.676
k	40	1.936	1.729	1.438	1.311	3.481
	100	1.806	1.626	1.396	1.300	2.241
Avg. n_i	80	2.008	1.756	1.481	1.334	3.294
	400	1.735	1.599	1.352	1.277	2.429
Predictor inter-corrs.	Low	1.931	1.666	1.479	1.345	2.712
	High	1.811	1.689	1.354	1.267	3.011
V_{mod}	.006	2.002	1.742	1.450	1.305	3.168
	.03	1.740	1.613	1.383	1.306	2.554
Incidence of M. Data	50%	1.496	1.387	1.285	1.190	2.158
	75%	2.246	1.968	1.548	1.421	3.565
M. Data Mechanism	sp-NMAR	2.017	1.677	1.514	1.317	2.814
	mp-NMAR	1.725	1.678	1.320	1.294	2.908

The lowest relative efficiencies for the estimation of β_0 (between 1.10 and 1.25)

and β_1 (1.20 to 1.30) came from cases with an average study sample size of 400, k = 100,

and only 50% missing data on the predictors. V_{mod} tended to be .03, but not always. The

lowest relative efficiencies for $\hat{\beta}_2$ and $\hat{\beta}_3$ (about 1.05 to 1.10) came from simulated meta-

analyses that had an average study sample size of 400, had large correlations among

predictors, and used mp-NMAR data. Finally, the lowest relative efficiencies for

estimation of τ (1.05 to 1.15) came from cases with an average study sample size of 400,

$V_{mod} = .03$, and 50% missing data on the predictors.

The only interactions that were substantively large in causing variation in the ratios

for τ were identical to those found for the MCAR/MAR data, and are reflected in Tables

5.32 - 5.35.

Table 5.32

MSE_{CC}/MSE_{MLE} Ratios for τ, p-NMAR Data (Incidence of Missing Data x Size of τ)		
		Ratio for τ
50% Incidence of Missing Data	$\tau = 0$	3.222
	$\tau = .005$	1.838
	$\tau = .02$	1.413
75% Incidence of Missing Data	$\tau = 0$	5.613
	$\tau = .005$	3.142
	$\tau = .02$	1.934

Table 5.33

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, p-NMAR Data		
(Average Study Sample Size x Size of τ)		
		Ratio for τ
Average $n_i = 80$	$\tau = 0$	4.857
	$\tau = .005$	3.518
	$\tau = .02$	1.911
Average $n_i = 400$	$\tau = 0$	4.544
	$\tau = .005$	1.625
	$\tau = .02$	1.477

Table 5.34

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, p-NMAR Data		
(Number of Studies x Size of τ)		
		Ratio for τ
$k = 40$	$\tau = 0$	5.502
	$\tau = .005$	3.072
	$\tau = .02$	1.869
$k = 100$	$\tau = 0$	3.333
	$\tau = .005$	1.908
	$\tau = .02$	1.482

Table 5.35

MSE_{CC}/MSE_{MLE} Ratios for τ, p-NMAR Data		
(Incidence of Missing Data x Number of Studies)		
		Ratio for τ
50% Incidence of Missing	$k = 40$	2.611
Data	$k = 100$	1.871
75% Incidence of Missing	$k = 40$	4.518
Data	$k = 100$	2.444

The patterns in the above tables are identical to those found in Tables 5.11 through

5.14 for the MCAR/MAR data. Estimation of τ is especially efficient for the ML method

when τ is 0 and there is either 75% missing data on the predictors, there is a small average

study sample size, or there is a small number of studies. Finally, ML estimation is

especially efficient when there is a small number of studies and a large amount of missing

data.

MSE_{CC} to MSE_{AC} Ratios

Table 5.36 summarizes the MSE ratios across all simulation conditions. As was the

case with the MCAR/MAR data, ratios were sometimes practically zero due to poor

estimation and extreme outliers in estimates of β. These cases came about in conditions in

which there was 75% missing data, $k = 40$, and there were high correlations among

predictors.

Table 5.36

MSE$_{CC}$/MSE$_{AC}$ Ratios, p-NMAR Data					
	β_0	β_1	β_2	β_3	τ
Mean	1.324	1.214	1.045	.924	1.011
Median	1.280	1.219	1.063	.969	1.061
Minimum	.037	.006	.005	.004	.240
Maximum	2.682	1.978	1.763	1.343	2.549

On average, available-case methods provide better estimates than complete-case methods for β_0, β_1, but little to no improvement in estimation for β_2, β_3, and τ. The results vary considerably across the values of the hyperparameters: Table 5.37 summarizes the MSE ratios by main effect. Table 5.38 summarizes the ratios across combinations of the different values of missing data incidence, k, and size of predictor intercorrelations. Just as for the MCAR/MAR data, the largest main effects and interactions were among and between these simulation hyperparameters.

The patterns of ratios are similar to those for the MCAR/MAR data. The pairwise estimates have the highest relative efficiency when the correlations among the variables are the lowest, and especially when the correlations among the predictors and between the predictors and the outcome are lowest. Except for β_2, the ratios are all closer to 1.00 than they were for the MCAR/MAR data.

95

Table 5.37

| MSE_{CC}/MSE_{AC} Ratios, Main Effects (p-NMAR Data) |
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.201	1.132	.969	.856	1.074
	.005	1.276	1.179	1.025	.903	1.124
	.02	1.495	1.330	1.142	1.014	1.315
k	40	1.344	1.100	.947	.831	1.309
	100	1.304	1.327	1.114	1.017	1.033
Avg. n_i	80	1.493	1.310	1.183	.970	1.376
	400	1.155	1.117	1.044	.878	.967
Predictor inter-corrs.	Low	1.464	1.338	1.159	1.026	1.191
	High	1.183	1.089	.931	.822	1.151
V_{mod}	.006	1.484	1.232	1.059	.914	1.245
	.03	1.163	1.195	1.031	.934	1.098
Incidence of M. Data	50%	1.141	1.135	1.042	.935	.898
	75%	1.506	1.292	1.049	.913	1.445
M. Data Mechanism	sp-NMAR	1.378	1.199	1.132	.918	1.154
	mp-NMAR	1.270	1.228	.959	.930	1.188

Table 5.38

MSE$_{CC}$/MSE$_{AC}$ Ratios, p-NMAR Data (Incidence of Missing Data x k x Predictor Intercorrelations)				β_0	β_1	β_2	β_3
50% Missing Data	$k = 40$		Low Corrs.	1.272	1.207	1.119	1.001
			High Corrs.	1.192	1.118	.989	.884
	$k = 100$		Low Corrs.	1.101	1.133	1.081	.969
			High Corrs.	1.000	1.084	.977	.887
75% Missing Data	$k = 40$		Low Corrs.	1.795	1.396	1.092	.947
			High Corrs.	1.117	.681	.590	.491
	$k = 100$		Low Corrs.	1.690	1.619	1.344	1.187
			High Corrs.	1.423	1.471	1.170	1.027

AC estimation of the slopes has the largest gains in efficiency over the CC estimates when $\tau = .02$, and in these conditions, AC estimation sometimes compares favorably with the maximum likelihood method, as shown in Table 5.39. Maximum-likelihood estimation of τ is on average 29% more efficient for these conditions, the estimation of the intercept 21% more efficient, but the estimation of the slopes is on average only 5 to 10 percent more efficient. In nine of the 192 conditions, at least one of the hyperparameters was marginally less efficiently estimated through ML estimation. In all of these conditions τ was .02, in eight, the data was sp-NMAR, and in seven, V_{mod} was .006.

Table 5.39

	β_0	β_1	β_2	β_3	τ
MSE$_{AC}$/MSE$_{MLE}$ Ratios for p-NMAR Data, $\tau = .02$ (Excluding k=40/75% Incidence of Missing Data)					
Mean	1.211	1.100	1.057	1.095	1.292
Median	1.160	1.102	1.070	1.104	1.279
Minimum	.976	1.025	.899	.921	1.095
Maximum	1.866	1.171	1.251	1.290	1.515

3. Results: o-NMAR Data

There were a total of 96 combinations of hyperparameters for which the missing-data mechanism was related to the population value of the outcome. Data of this type is referred to as o-NMAR data.

Bias in ML Estimation of β

Maximum-likelihood theory suggests that when the missing-data mechanism is a function of the observed values of the outcome, maximum-likelihood estimates will be asymptotically unbiased. However, as noted in Chapter 3, Little (1992)'s ML estimates when the missing-data mechanism was a function of the values of the outcome seemed to be biased. Also, as described in Chapter 4, in this investigation the missing-data mechanism was related to the *population* value of the outcome. The population value is not known to a meta-analyst, however; only the sample values are known.

For the 96 conditions, there were 85 biased estimates of β_0, 80 biased estimates of β_1, 43 biased estimates of β_2, and 62 biased estimates of β_3. Table 5.40 shows the average MSE/variance ratio across all conditions for each slope. Except for four outliers (all of which have in common $\tau = 0$ or $.005$, $V_{mod} = .03$, low correlations among predictors, and average $n_i = 400$), all ratios for β_2 and β_3 are below 1.12, implying that most of the time the size of the bias is small relative to the sampling error in the estimates. The ratios for β_0 and for β_1 have larger ratios under some conditions. A breakdown of the ratios and biases across conditions is in Table 5.41.

Table 5.40

	β_0	β_1	β_2	β_3
Maximum-Likelihood MSE/Variance Ratios for β (o-NMAR Data)				
Mean	1.509	1.063	1.025	1.021
Median	1.189	1.034	1.005	1.013
Minimum	1.000	1.000	1.000	1.000
Maximum	4.656	1.356	1.336	1.235

The MSE/variance ratios in Table 5.41 are largest for $\tau = 0$; on average, over half the mean-squared error in the estimation of β_0 is caused by bias in the estimation of β_0. However, on average, the bias is slight, and substantively insignificant relative to the metric of standardized mean differences. The ratios for β_1 are less extreme; they are the highest for $\tau = 0$, but the average biases are largest for $\tau = .02$. However, for this condition the average ratio is only 1.02, indicating that the bias is generally insignificant relative to the sampling error in $\hat{\beta}_1$.

100

Table 5.41

<table>
<tr><th colspan="6">Maximum-Likelihood MSE/Variance Ratios and Biases
for β_0 and β_1, Main Effects (o-NMAR Data)</th></tr>
<tr><th>Parameter</th><th>Parameter
Value</th><th>Ratio for β_0</th><th>Bias in $\hat{\beta}_0$</th><th>Ratio for β_1</th><th>Bias in $\hat{\beta}_1$</th></tr>
<tr><td>τ</td><td>0</td><td>2.057</td><td>-.0281</td><td>1.127</td><td>-.0104</td></tr>
<tr><td></td><td>.005</td><td>1.380</td><td>-.0177</td><td>1.043</td><td>-.0623</td></tr>
<tr><td></td><td>.02</td><td>1.089</td><td>-.0049</td><td>1.020</td><td>-.0724</td></tr>
<tr><td>k</td><td>40</td><td>1.205</td><td>-.0153</td><td>1.029</td><td>-.0481</td></tr>
<tr><td></td><td>100</td><td>1.813</td><td>-.0185</td><td>1.098</td><td>-.0487</td></tr>
<tr><td>Avg. n_i</td><td>80</td><td>1.391</td><td>-.0207</td><td>1.059</td><td>-.0448</td></tr>
<tr><td></td><td>400</td><td>1.627</td><td>-.0131</td><td>1.068</td><td>-.0520</td></tr>
<tr><td>Predictor
inter-corrs.</td><td>Low</td><td>1.664</td><td>-.0221</td><td>1.065</td><td>-.0622</td></tr>
<tr><td></td><td>High</td><td>1.354</td><td>-.0117</td><td>1.062</td><td>-.0345</td></tr>
<tr><td>V_{mod}</td><td>.006</td><td>1.271</td><td>-.0080</td><td>1.045</td><td>-.0608</td></tr>
<tr><td></td><td>.03</td><td>1.747</td><td>-.0259</td><td>1.082</td><td>-.0359</td></tr>
<tr><td>Incidence of
M. Data</td><td>50%</td><td>1.435</td><td>-.0139</td><td>1.051</td><td>-.0362</td></tr>
<tr><td></td><td>75%</td><td>1.583</td><td>-.0199</td><td>1.076</td><td>-.0606</td></tr>
</table>

Bias in ML Estimation of τ

An ANOVA was conducted to determine the extent to which the seven simulation patterns were related to variation in the size of the bias in the estimates of τ. All seven simulation hyperparameters and two-way interactions between those hyperparameters were significant; however, the only substantive main effects came from missing data

101

pattern, number of studies and the value of τ. Table 5.42 shows the average ratios and biases across combinations of these hyperparameters. There was a substantively important interaction effect between k and τ: large values of k lessen the negative bias in $\hat{\tau}$. There was one other substantively large interaction, shown in Table 5.43, between average study sample size and τ: large average study sample sizes both decrease the positive bias in $\hat{\tau}$ for $\tau = 0$ and decrease the negative bias in $\hat{\tau}$ for $\tau = .02$. These results are very similar to those of the MCAR/MAR data; in general, biases are quite small, especially when $k = 100$.

Table 5.42

Maximum-Likelihood MSE/Variance Ratios and Biases for τ (o-NMAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	k = 40	$\tau = 0$	1.173	.0015
		$\tau = .005$	1.151	-.0016
		$\tau = .02$	1.302	-.0067
	k = 100	$\tau = 0$	1.385	.0014
		$\tau = .005$	1.042	-.0005
		$\tau = .02$	1.063	-.0024
75% Incidence of Missing Data	k = 40	$\tau = 0$	1.150	.0014
		$\tau = .005$	1.349	-.0021
		$\tau = .02$	1.754	-.0105
	k = 100	$\tau = 0$	1.353	.0016
		$\tau = .005$	1.098	-.0008
		$\tau = .02$	1.189	-.0044

Table 5.43

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, o-NMAR Data				
(Average Study Sample Size x Population Variance)				
			τ Ratio	Bias in $\hat{\tau}$
Average n_i = 80		τ = 0	1.252	.0024
		τ = .005	1.030	-.0008
		τ = .02	1.323	-.0076
Average n_i = 400		τ = 0	1.278	.0006
		τ = .005	1.290	-.0017
		τ = .02	1.332	-.0045

Bias in AC Estimation of β

Significant biases in the AC estimators were more frequent for the o-NMAR data than they were for any other type of data. For the 96 conditions, estimates of β_0 were biased 95 times, estimates of β_1 were biased 90 times, estimates of β_2 were biased 86 times, and estimates of β_3, 84 times. A summary of the MSE/variance ratios to determine the extent of the bias relative to the sampling error of the estimates is in Table 5.44. These ratios are higher than they were for the other data types in previous sections. A summary of the average biases is in Table 5.45, and the main effects of the simulation hyperparameters on ratio and bias are shown in Tables 5.46 and 5.47. However, as can be seen in Table 5.45, there were many outliers with regard to bias for β_1, β_2, and β_3, which makes interpretation of factors that cause bias in the estimators of those parameters difficult. There were fewer stronger outliers for β_0; it can be concluded from Table 5.46

that bias is lower and the size of the bias relative to sampling error is lower for the high

values of τ and low values of V_{mod}. These findings mimic those found for the

MCAR/MAR and p-NMAR data.

Table 5.44

Available-Case MSE/Variance Ratios for β (o-NMAR Data)				
	β_0 Ratio	β_1 Ratio	β_2 Ratio	β_3 Ratio
Mean	2.081	1.125	1.095	1.085
Median	1.425	1.058	1.034	1.032
Minimum	1.000	1.000	1.000	1.000
Maximum	8.370	1.639	1.821	1.529

Table 5.45

Available-Case Biases for β (o-NMAR Data)				
	Bias in $\hat{\beta}_0$	Bias in $\hat{\beta}_1$	Bias in $\hat{\beta}_2$	Bias in $\hat{\beta}_3$
Mean	-.0400	.1083	.1333	.0910
Median	-.0360	.0985	.1161	.0730
Lowest	-.1474	-.0322	-.0530	-1.085
Highest	-.0026	.5481	.6663	.5701

Table 5.46

Available-Case MSE/Variance Ratios and Biases for β_0 and β_1, Main Effects (o-NMAR Data)					
Parameter	Parameter Value	Ratio for β_0	Bias in $\hat{\beta}_0$	Ratio for β_1	Bias in $\hat{\beta}_1$
τ	0	2.658	-.0540	1.149	.1027
	.005	2.091	-.0390	1.124	.1003
	.02	1.493	-.0260	1.101	.1229
k	40	1.396	-.0410	1.037	.1087
	100	2.766	-.0390	1.213	.1078
Avg. n_i	80	1.581	-.0380	1.087	.1167
	400	2.580	-.0420	1.162	.0998
Predictor inter-corrs.	Low	2.173	-.0440	1.010	.0752
	High	1.989	-.0360	1.152	.1414
V_{mod}	.006	1.320	-.0190	1.058	.1242
	.03	2.841	-.0610	1.192	.0923
Incidence of M. Data	50%	2.058	-.0300	1.123	.0962
	75%	2.103	-.0500	1.125	.1204

Table 5.47

Available-Case MSE/Variance Ratios and Biases for β_2 and β_3, Main Effects (o-NMAR Data)					
Parameter	Parameter Value	Ratio for β_2	Bias in $\hat{\beta}_2$	Ratio for β_3	Bias in $\hat{\beta}_3$
τ	0	1.156	.1753	1.063	.0562
	.005	1.092	.1135	1.073	-.1030
	.02	1.036	.1112	1.118	-.2264
k	40	1.048	.1679	1.032	-.0800
	100	1.142	.0988	1.137	-.1019
Avg. n_i	80	1.049	.1280	1.060	-.1227
	400	1.141	.1387	1.109	-.0590
Predictor inter-corrs.	Low	1.106	.1147	1.066	.0313
	High	1.084	.1520	1.103	-.2140
V_{mod}	.006	1.031	.1251	1.071	-.1710
	.03	1.158	.1416	1.098	-.0110
Incidence of Missing Data	50%	1.108	.1727	1.094	-.0920
	75%	1.082	.0940	1.077	-.0900

<u>Bias in AC Estimation of τ</u>

Table 5.48 summarizes the sizes of the average bias in $\hat{\tau}$ for the hyperparameters most strongly related to the size of the bias. Unlike for the MCAR/MAR and p-NMAR data, the percentage of missing data was relatively unimportant.The hyperparameter V_{mod} was the third most important hyperparameter, having roughly the same size effect as

106

variation in τ.

Table 5.48

Available-Case MSE/Variance Ratios and Biases for τ (o-NMAR Data)				
			Ratio for τ	Bias in $\hat{\tau}$
$V_{mod} = .006$	Average $n_i = 80$	$\tau = 0$	1.371	.0075
		$\tau = .005$	1.153	.0057
		$\tau = .02$	1.017	.0018
	Average $n_i = $ 400	$\tau = 0$	1.333	.0026
		$\tau = .005$	1.031	.0011
		$\tau = .02$	1.018	-.0004
$V_{mod} = .03$	Average $n_i = 80$	$\tau = 0$	1.411	.0091
		$\tau = .005$	1.194	.0080
		$\tau = .02$	1.079	.0059
	Average $n_i = $ 400	$\tau = 0$	1.342	.0032
		$\tau = .005$	1.073	.0022
		$\tau = .02$	1.056	.0018

Average study sample size was the most important predictor, as was the case with the MCAR/MAR and p-NMAR data. When the average study sample size was only 80, there were very large biases in $\hat{\tau}$ for $\tau = 0$ and $\tau = .005$. This bias was even higher when $V_{mod} = .03$. Biases were generally small for estimates coming from simulated meta-analyses with an average study sample size of 400, especially when the predictors explained only a small amount of variation in the outcome ($V_{mod} = .006$).

107

Out of the 96 conditions, there were 64 instances of a significant bias for β_0, 46 instances of significant bias for β_1, 52 instances of significant bias for β_2, and 63 instances of significant bias for β_3. A summary of the overall MSE/Variance ratios are in Table 5.49 and a summary of the biases are in Table 5.50. For the MCAR/MAR and p-NMAR data, the ratios and biases were typically very small; this is not the case for the o-NMAR data. Estimation of the intercept in particular can be very biased, and that bias can be large relative to the sampling error of the estimator. This is expected given that the true value of the outcome was strongly related to the missing-data mechanism. All ratios are higher than they were for the MCAR/MAR and p-NMAR data.

Table 5.49

Complete-Case MSE/Variance Ratios for β (o-NMAR Data)				
	Ratio for β_0	Ratio for β_1	Ratio for β_2	Ratio for β_3
Mean	2.393	1.020	1.041	1.072
Median	1.357	1.006	1.012	1.025
Minimum	1.000	1.000	1.000	1.000
Maximum	11.458	1.228	1.421	1.711

Table 5.50

Complete-Case Biases for β (o-NMAR Data)				
	Bias in $\hat{\beta}_0$	Bias in $\hat{\beta}_1$	Bias in $\hat{\beta}_2$	Bias in $\hat{\beta}_3$
Mean	.0509	-.0484	-.0703	-.1111
Median	.0402	-.0397	-.0534	-.1088
Lowest	-.0100	-.1700	-.2500	-.3100
Highest	.1500	.0700	.1200	.0300

Bias in CC Estimation of τ

Patterns in the bias in the CC $\hat{\tau}$s are similar to those found for the p-NMAR data. Table 5.51 summarizes the sizes of the bias over τ, k, incidence of missing data, and average n_i. Biases in the estimation of low values of τ were very large, sometimes over .01, in cases where the average study sample size was small. There were substantial interaction effects between average study sample size and both incidence of missing data and number of studies. A high proportion of missing data or low value for k exacerbated the bias in the estimation of τ.

Table 5.51

Complete-Case MSE/Variance Ratios and Biases for τ (o-NMAR Data)		Ratio for τ	Bias in $\hat{\tau}$
$\tau = 0$	Average $n_i = 80$	1.400	.0119
	Average $n_i = 400$	1.383	.0022
$\tau = .005$	Average $n_i = 80$	1.183	.0094
	Average $n_i = 400$	1.025	.0005
$\tau = .02$	Average $n_i = 80$	1.020	.0030
	Average $n_i = 400$	1.037	-.0052
$k = 40$	Average $n_i = 80$	1.193	.0117
	Average $n_i = 400$	1.173	-.0002
$k = 100$	Average $n_i = 80$	1.209	.0045
	Average $n_i = 400$	1.347	-.0015
50% Incidence of Missing Data	Average $n_i = 80$	1.205	.0052
	Average $n_i = 400$	1.285	-.0011
75% Incidence of Missing Data	Average $n_i = 80$	1.197	.0110
	Average $n_i = 400$	1.234	-.0005

110

Table 5.52 summarizes the MSE ratios across all simulation conditions. The results for β_1, β_2, β_3, and τ are similar to those for the MCAR/MAR and NMAR data, while the results for β_0 are more varied. Estimation of β_0 is usually over twice as efficient, but there are instances in which CC estimation is more efficient than ML estimation.

Table 5.52

MSE$_{CC}$/MSE$_{ML}$ Ratios (o-NMAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	2.332	1.612	1.261	1.248	2.505
Median	2.204	1.490	1.230	1.210	2.161
Minimum	.672	1.041	.868	.845	.924
Maximum	5.620	2.624	1.988	1.829	8.516

The relative efficiency varied depending on the simulation conditions. Table 5.53 shows the main effects of each of the seven simulation hyperparameters. There were no substantively interesting interaction effects among the simulation hyperparameters for any β or τ. The effects are very similar to those for the MCAR/MAR and p-NMAR data: smaller values of τ are related to higher relative efficiencies for ML estimation of the slopes and τ, as are smaller average sample sizes and larger proportions of missing data. The primary difference in the o-NMAR data is that the effect of the size of τ on relative efficiency is reversed for estimation of β_0.

Table 5.53

MSE$_{CC}$/MSE$_{MLE}$ Ratios, Main Effects (o-NMAR Data)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.669	1.689	1.416	1.401	3.766
	.005	2.178	1.616	1.258	1.242	2.275
	.02	3.148	1.531	1.111	1.110	1.475
k	40	2.148	1.671	1.290	1.248	3.028
	100	2.515	1.553	1.232	1.247	1.983
Avg. n_i	80	2.288	1.727	1.336	1.289	2.969
	400	2.375	1.497	1.187	1.206	2.041
Predictor inter-corrs.	Low	2.164	1.635	1.306	1.287	2.339
	High	2.499	1.589	1.217	1.208	2.672
V_{mod}	.006	2.675	1.687	1.290	1.248	2.827
	.03	1.988	1.537	1.232	1.247	2.184
Incidence of M. Data	50%	1.978	1.349	1.171	1.168	1.922
	75%	2.685	1.876	1.352	1.327	3.088

ML estimation is over three times as efficient as CC estimation when τ =.02, but only about one-and-a-half times as efficient when τ = 0. The reason for this stems from the fact that CC estimation of data that is NMAR because of variation in the outcome is worse the more unexplained variation there is. When τ = 0, the only variation in the population value of the outcome stems from variation on the predictors, whose values are known. When τ = .02, most variation in the outcome stems from random error.
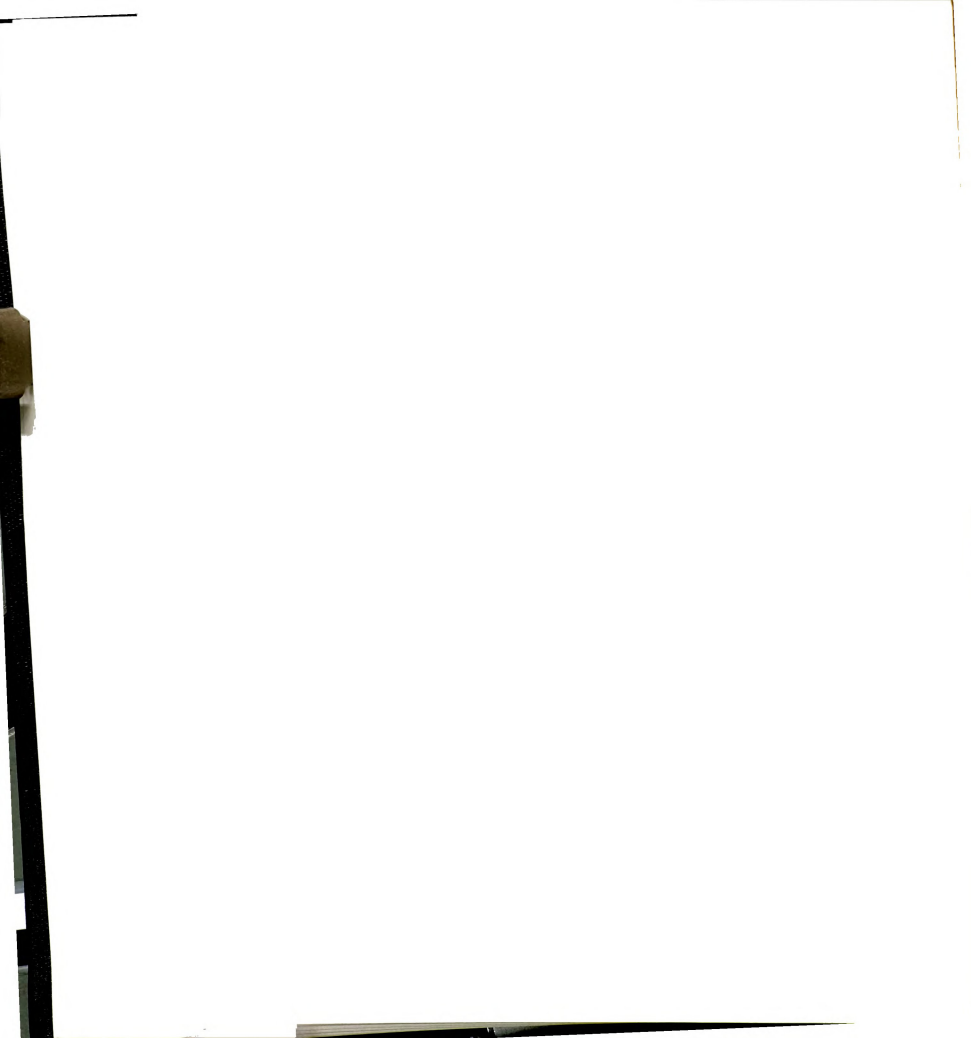
The lowest relative efficiencies for the estimation of β_0 (there were six conditions between .67 and 1.20, four under 1.00) had much in common; all had $\tau = 0$, as expected. For all but one of the six conditions, $k = 100$, $V_{mod} = .03$, average $n_i = 400$, there was 75% missing data, and there were low correlations between the predictors.

The lower relative efficiencies of β_1 (1.04 to 1.25) came from cases with average $n_i = 400$ and 75% missing data. There were eleven conditions for which the relative efficiency of either β_2 or β_3 (or both) was less than 1.00. In these conditions, it was generally the case that $\tau = .02$, $V_{mod} = .006$, the average $n_i = 400$, and there were high correlations between predictors. Finally, there were only two conditions for which the relative efficiency of τ was less than zero; all of the lower efficiencies for τ (i.e., below 1.15) tended to come from conditions in which τ was .01 or .02 and there was 75% missing data on the predictors.

MSE$_{CC}$ to MSE$_{AC}$ Ratios

Table 5.54 summarizes the MSE ratios across all simulation conditions.

Table 5.54

MSE$_{CC}$/MSE$_{AC}$ Ratios (o-NMAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.564	1.158	.913	.883	1.073
Median	1.466	1.164	.965	.932	.995
Minimum	.101	.114	.059	.059	.245
Maximum	5.277	2.175	1.355	1.290	2.382

On average, available-case methods provide much better estimates than complete-case methods for β_0, slightly better estimates for β_1, and τ, and slightly inferior estimates for β_2 and β_3. The results vary considerably across the values of the hyperparameters; Table 5.55 summarizes the MSE ratios by main effect

The patterns of ratios are similar to those for the MCAR/MAR data. The pairwise estimates have the highest relative efficiency when the correlations among the variables are the lowest, and especially when the correlations among the predictors and between the predictors and the outcome are lowest.

114

Table 5.55

MSE$_{CC}$/MSE$_{AC}$ Ratios, Main Effects (o-NMAR Data)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	.931	1.058	.821	.818	1.068
	.005	1.374	1.171	.946	.910	1.002
	.02	2.387	1.243	.971	.922	1.150
k	40	1.365	1.065	.818	.783	1.205
	100	1.763	1.250	1.007	.983	.941
Avg. n_i	80	1.610	1.293	.995	.949	1.299
	400	1.518	1.022	.830	.817	.848
Predictor inter-corrs.	Low	1.597	1.242	.956	.941	1.054
	High	1.531	1.073	.879	.825	1.092
V_{mod}	.006	2.002	1.275	.970	.923	1.171
	.03	1.126	1.040	.855	.844	.976
Incidence of M. Data	50%	1.436	1.089	.930	.909	.818
	75%	1.692	1.226	.895	.857	1.328

AC estimation of the intercept is far better for $\tau = .02$ and $V_{mod} = .006$. For other values of τ or for $V_{mod} = .03$, AC estimation of β_0 is only marginally better than CC estimation. There are no hyperparameters that have as large an effect on the relative efficiency for estimation of β_1; however, there are moderate-sized effects for all of the simulation hyperparameters. Similarly, there are no hyperparameters that have a large effect on the relative efficiencies for estimation of β_2 or β_3, but there are moderate effects

for k, average $n_{i.}$, V_{mod} and size of predictor intercorrelations. There were substantively interesting interaction effects for the slopes between k and incidence of missing data, as shown in Table 5.56.

Table 5.56

MSE_{CC}/MSE_{AC} Ratios for NMAR Data					
(Incidence of Missing Data x k)					
		β_0	β_1	β_2	β_3
50% Incidence	$k = 40$	1.373	1.120	.942	.903
of Missing Data	$k = 100$	1.499	1.059	.918	.915
75% Incidence	$k = 40$	1.357	1.010	.694	.663
of Missing Data	$k = 100$	2.027	1.442	1.096	1.052

Available-case estimation of β_0 and β_1 is especially good relative to complete-case estimation when there are a large amount of studies and a large amount of missing data, while AC estimation of β_2 and β_3 is especially poor when there are a small amount of studies and a large amount of missing data.

While in general the results of the above sections suggest that for o-NMAR data, ML estimation is superior to AC estimation, there is a subset of conditions for which AC estimation is almost as efficient as ML estimation.

Table 5.57

MSE$_{AC}$/MSE$_{MLE}$ Ratios for o-NMAR Data, $\tau = .02$ (Excluding k=40/75% Missing Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.333	1.141	1.061	1.126	1.312
Median	1.196	1.121	1.046	1.088	1.305
Minimum	.980	1.034	.894	.905	.990
Maximum	2.079	1.405	1.510	1.668	1.570

For a low incidence of missing data or a large number of studies, AC estimation of

the slopes is on average only marginally more efficient than ML estimation. In some cases

(all of which have in common a k of 100) estimation of some slopes is slightly worse using

the ML method.

4. Estimation of the Population Mean

One last parameter worth considering is the population mean of all of the effect sizes. Whether this parameter is of interest depends on the purposes of the meta-analysis. Given estimates of β_0, β_1, β_2, and β_3, the meta-analyst can determine the expected value for any typical study, given values of the study predictors X_1, X_2, and X_3. Thus, for example, the mean effect size (measuring efficacy of treatment) might not be of interest in a meta-analysis where the first predictor is *length of treatment* (in weeks), the 2nd predictor is *average contact time per week,* and the 3rd predictor is the *age of the subjects.* Average treatment efficacy may be of less interest than how successful the treatment might be given specific values of the predictors (i.e., specific lengths of programs, ages of youths, and average contact time per week).

On the other hand, there will be times when the meta-analyst is interested in the average effect, regardless of what the values of the predictors might be. For instance, in the example above, it is certainly possible that the meta-analyst would want to know whether the average treatment received by a juvenile delinquent today tends to help, irrespective of what the treatment might be.

Because all predictors in the simulation had a population mean of zero (note: this is different than saying that all were "mean-centered"), estimation of the mean for all three methods is essentially identical to the estimation of β_0. As noted early in the first section of this chapter, estimation of β_0 was essentially unbiased for all three estimation procedures (ML estimation, AC estimation, and CC estimation). This was theoretically expected.

118

<u>Bias in Mean Estimation: MAR Data</u>

Theoretically, there should be no bias in the ML estimator of the mean. However, given the missing-data mechanism in question, there should be a positive bias in the CC estimator. There is no theory to suggest that the AC estimator should be unbiased (though there is also no theory to suggest whether any existing bias should be either positive or negative.) These expectations were borne out in the analyses. Only one of the 96 combinations of hyperparameters examined for the MAR data showed a statistically significant bias ($p<.01$) in ML estimation of the mean. However, all of the 96 conditions led to bias in the CC estimation of the mean, and 70 of the 96 conditions led to bias in the AC estimation of the mean.

As shown in Table 5.58, the size of the bias relative to the standard error of estimate (examined by considering the empirical MSE/variance ratios) was minimal for ML estimation, and often trivial for AC estimation. However, for CC estimation the ratios were often considerable, relative to the size of the mean-squared error of estimation.

Table 5.58

MSE/Variance Ratios for Estimation of the Mean (MAR Data)			
	ML	AC	CC
Mean	1.001	1.063	2.634
Median	1.001	1.015	2.137
Minimum	1.000	1.000	1.118
Maximum	1.010	3.706	10.035

While the ratios above seem large, the actual size of the bias for the CC estimator of the mean ranged from small to moderate in size. This is easily explained: the mean was usually measured with extreme precision. The average CC bias across the conditions ranged from +.02 to +.12, with an average of +.06. In the effect size metric, the accepted guidelines are that a small effect size is .2, a medium effect size is .5, and a large effect size is .8 (Cohen, 1988). The importance of the size of this bias will depend on what the mean effect size actually is. The difference in effect size between an estimate of .62 and an estimate of .50 is noticeable, but will perhaps be substantively unimportant to some researchers. On the other hand, the difference in effect size between an estimate of .20 and an estimate .08 might be very important. Conceivably, it could mean the difference between wide-scale implementation of a treatment, and sending it back to the drawing board.

This direction of bias was expected, as the missing-data mechanism was such that studies were more likely to be missing when they had lower values on the 1st predictor. Thus, the observed studies were more likely to have higher values on the 1st predictor. Given the positive relationship between the 1st predictor and the outcome, the complete-case studies were more likely to have higher effect sizes than the population mean.

Bias in Mean Estimation: p-NMAR Data

Theoretically, there should be bias in all three estimators of the mean when there is this type of missing data. These expectations were borne out in the analysis, though ML estimation clearly excelled. There were two types of p-NMAR data: sp-NMAR data

120

(strongly NMAR) and mp-NMAR data (moderately NMAR). Only 16 of the 96 combinations of hyperparameters examined for the sp-NMAR data led to a statistically significant bias in the ML estimator of the mean. However, all 96 conditions led to statistically significant biases in both the CC and AC estimators of the mean.

As expected, the biases were somewhat less severe for the mp-NMAR data. Only three of the 96 combinations of hyperparameters examined for the mp-NMAR data led to a statistically significant bias in the ML estimator of the mean. Again, the CC estimator of the mean was biased across all 96 conditions. However, the AC estimator of the mean was biased in only 86 of those conditions.

As shown in Table 5.59 and Table 5.60, the size of the bias relative to the standard error of estimate (examined by considering the empirical MSE/variance ratios) was minimal for EM estimation, and often trivial for AC estimation. However, for CC estimation it was often considerable, relative to the size of the mean-squared error of estimation.

Table 5.59

MSE/Variance Ratios for Estimation of the Mean (sp-MAR Data)			
	EM	AC	CC
Mean	1.004	1.347	3.491
Median	1.003	1.190	2.743
Minimum	1.000	1.003	1.175
Maximum	1.035	2.902	15.481

Table 5.60

MSE/Variance Ratios for Estimation of the Mean (np-MAR Data)			
	EM	AC	CC
Mean	1.001	1.084	1.589
Median	1.001	1.045	1.396
Minimum	1.000	1.000	1.024
Maximum	1.010	1.414	4.091

These ratios are similar to those found in Table 5.58. The average CC bias across the sp-NMAR conditions ranged from +.03 to +.14, with an average of +.07, while the AC biases for the same set of conditions ranged from -.05 to -.01, with an average of -.02. The average CC bias across the mp-NMAR conditions ranged from +.01 to +.07, with an average of +.04, while the AC biases for those conditions ranged from -.03 to +.01, with an average of -.01. Relative to the effect size metric, only the biases for the CC estimator are likely to be high enough to lead to substantively incorrect conclusions.

Bias in Mean Estimation: o-NMAR Data

As with the other NMAR data, there should theoretically be bias in all three estimators of the mean. These expectations were borne out in the analysis – again, ML estimation was clearly superior. In 31 of the 96 combinations of hyperparameters examined for the o-NMAR data, there was a statistically significant bias ($p<.01$) in the ML estimation of the mean. Again, all of the 96 conditions led to bias in the CC estimation of

122

the mean, and most (89 of the 96 conditions) led to bias in the AC estimation of the mean.

As shown in Table 5.61, the sizes of the biases relative to the standard error of estimate were greater for this condition than for any of the others. The bias was still minimal for ML estimation. However, it was small to moderate for AC estimation, and often very large for CC estimation.

Table 5.61

MSE/Variance Ratios for Estimation of the Mean (o-NMAR Data)			
	ML	AC	CC
Mean	1.007	1.381	8.816
Median	1.003	1.174	7.763
Minimum	1.000	1.000	1.597
Maximum	1.065	3.698	28.31

The size of the bias ranged from +.05 to +.23 for the CC estimator of the mean, with an average bias of +.12. The bias of the AC estimator of the mean ranged from -.07 to zero, with an average bias of -.02. Again, the bias in the CC estimator is considerable enough relative to the effect size metric that it might lead researchers to conclusions that are, substantively speaking, far from correct.

5. Results: Dichotomous Predictor with Missing Data

The EM maximum-likelihood estimation method derived in Chapter III assumes that each predictor is normally distributed. However, it is generally accepted that EM estimation does not suffer when there are completely observed dichotomous variables as predictors. Nevertheless, there will be many times in meta-analysis when there are dichotomous variables of import that have some degree of missing data. One way to handle this problem is to derive a new EM estimation method for use in meta-analysis, one which combines continuous predictors and those with a multinomial distribution. However, that is a complex task, and the assumption of normality in maximum-likelihood estimation is often not that stringent. Thus, before pursuing it, it is worthwhile to investigate the possibility that EM estimation with dichotomous predictors with missing data is comparable to EM estimation with continuous predictors with regard to bias and improvement in MSE over CC estimation.

Computing time was not available to investigate the effect of having a dichotomous predictor or multiple dichotomous predictors in EM estimation of the meta-analytic model. The task is worthy of a large simulation study in itself; not only would the seven variables already varied in the present investigation be considered as key simulation hyperparameters, but consideration would have to be given to the proportions involved for the dichotomous variables. There might be different results for a dichotomous variable for which the division between the two groups is 50%/50% and a variable for which the division between the two groups is 90%/10%. There is also the problem that depending on the missing-data percentages and number of studies in the simulated meta-analyses, there

would be some simulations for which there could be no complete-case analysis using all variables (because in the complete data, all values for that variable are identical). This would especially be a risk when there are small amounts of studies, large percentages of missing data, a large population proportion in one group or the other (such as a 90%/10% split), or a missing-data mechanism strongly dependent on the value of one of the predictors or the outcome.

Because of these concerns, only a preliminary investigation into the question was conducted. To avoid the problems described in the previous paragraph, the meta-analyses were limited to instances in which $k = 100$ and the missing-data mechanism was MCAR. Within these constraints, values of V_{mod}, average n_i, τ, strength of intercorrelations between predictors, and missing data pattern were varied, leading to 48 combinations of hyperparameters. In these simulated meta-analyses, the third variable was treated as dichotomous; it was made dichotomous by following the procedure for generating data described in Chapter IV; once the values for the third variable were generated, each value was changed to -1.0 if the value was negative and to +1.0 if the value was positive. Obviously, in future investigation of this issue different values of k, different missing-data mechanisms, and different population proportions for the dichotomous variable should be considered.

<u>Bias in ML Estimation of β</u>

It was expected that except for the slope for the third predictor, the patterns of bias would be what they were in Section 1 of this Chapter. This indeed was the case; in the 48 conditions, there are 4 instances of bias for the intercept, 2 instances for β_1, 4 instances for β_2, and 12 instances for β_3 ($p<.01$). The significant biases occurred more often when the average study sample size was 400 and the missing data percentage was 75%.

Table 5.62 contains the average, median, minimum, and maximum empirical MSE/variance ratios. A histogram of MSE/variance ratios for β_3 is in Figure 5.4. The table and the figure show that even though there are more significant biases for estimation of β_3, the size of the biases remain small relative to the amount of sampling error in the estimators.

Table 5.62

Maximum-Likelihood MSE/Variance Ratios for β (MCAR Data w/Dichotomous Predictor)				
	β_0	β_1	β_2	β_3
Mean	1.002	1.002	1.004	1.006
Median	1.001	1.001	1.002	1.003
Minimum	1.000	1.000	1.000	1.000
Maximum	1.010	1.015	1.033	1.025

Figure 5.4
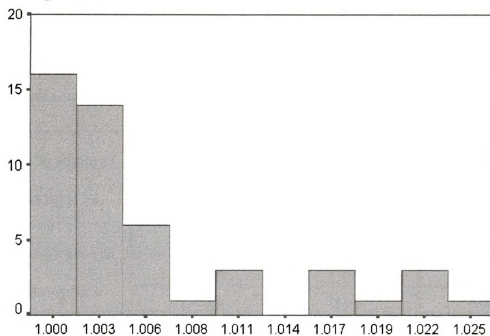
Ratio of MSE to Variance for β_3 (w/Dichotomous Predictor)

Bias in ML Estimation of τ

Table 5.63 shows the ratios and biases for the ML estimations across the different values of these simulation hyperparameters. The average ratio across all conditions was 1.260.
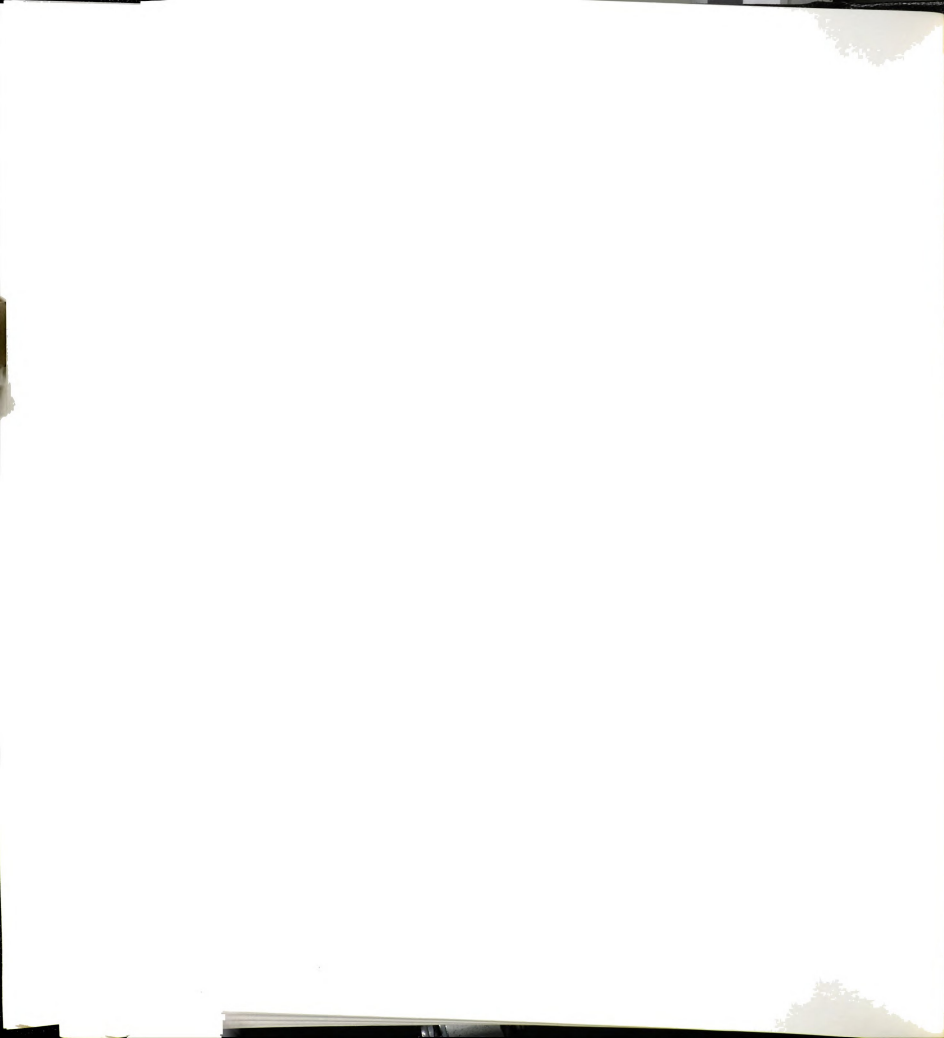
Table 5.63

Maximum-Likelihood MSE/Variance Ratios for τ (MCAR Data w/Dichotomous Predictor)			
		Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	$\tau = 0$	1.364	.0017
	$\tau = .005$	1.152	-.0012
	$\tau = .02$	1.151	-.0035
75% Incidence of Missing Data	$\tau = 0$	1.342	.0012
	$\tau = .005$	1.260	-.0017
	$\tau = .02$	1.293	-.0055

These results are almost identical to those for the MCAR/MAR data for $k = 100$ in section 1. of this chapter.

Bias in AC Estimation of β

Of the 48 combinations of simulation hyperparameters, there were 2 conditions for which there was a significant bias in the estimation of β_0, 4 conditions for estimation of β_1, 8 conditions for estimation of β_2, and 14 conditions for β_3. The proportion of cases for which there were significant biases for each parameter are roughly equivalent to those for continuous MCAR data, as expected given that there is no assumption in AC estimation that the predictors have continuous distributions. Table 5.64 demonstrates that the amount of bias in the AC estimators was similar to that in the MCAR data with continuous predictors: very small relative to the size of the variance of the estimates, even

128

in those few cases where the bias was statistically significant.

Table 5.64

Available-Case MSE/Variance Ratios for β (MCAR Data w/Dichotomous Predictor)				
	β_0	β_1	β_2	β_3
50% M.D. Incidence	1.001	1.001	1.001	1.002
75% M.D. Incidence	1.001	1.002	1.004	1.005

Bias in AC Estimation of τ

Table 5.65 summarizes the sizes of the average bias in $\hat{\tau}$ for the hyperparameters that were most strongly related to the size of bias. The hyperparameters selected – missing data incidence, average study sample size, and value of τ – are the same as for the MCAR data with continuous predictors. The findings are similar to those for the MCAR data with continuous predictors, as well. On average the biases are substantively small except for τ = 0 or .005 and average n_i = 80. Biases greater than zero are not unexpected for these conditions due to the fact that all $\hat{\tau}$ < 0 were set to 0 before calculations commenced.

129

Table 5.65

Available-Case MSE/Variance Ratios and Biases for τ				
(MCAR Data w/Dichotomous Predictor)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.436	.0066
		$\tau = .005$	1.147	.0047
		$\tau = .02$	1.009	.0015
	Average $n_i = 400$	$\tau = 0$	1.354	.0020
		$\tau = .005$	1.005	.0003
		$\tau = .02$	1.006	-.0006
75% Incidence of Missing Data	Average $n_i = 80$	$\tau = 0$	1.392	.0061
		$\tau = .005$	1.119	.0042
		$\tau = .02$	1.004	-.0002
	Average $n_i = 400$	$\tau = 0$	1.328	.0020
		$\tau = .005$	1.004	.0003
		$\tau = .02$	1.026	-.0014

Bias in CC Estimation of β and τ

As expected by estimation theory, there was little bias in the complete-case

estimators of β for the MCAR data; of the 48 conditions, there were no instances of

statistically significant bias for β_1 and only one instance for β_0, β_2, and β_3.

Bias in $\hat{\tau}$ was strong, as it was in the MCAR data with continuous predictors, as

can be seen in Table 5.66. The bias is especially strong for $\tau = 0$ or .005 and low average

sample size and large amounts of missing data.

Table 5.66

Complete-Case MSE/Variance Ratios and Biases for τ (MCAR Data w/Dichotomous Predictor)				
			Ratio for τ	Bias in $\hat{\tau}$
50% Incidence of Missing Data	Average n_i = 80	$\tau = 0$	1.428	.0058
		$\tau = .005$	1.153	.0044
		$\tau = .02$	1.011	.0017
	Average n_i = 400	$\tau = 0$	1.412	.0010
		$\tau = .005$	1.002	.0001
		$\tau = .02$	1.001	.0002
75% Incidence of Missing Data	Average n_i = 80	$\tau = 0$	1.395	.0093
		$\tau = .005$	1.184	.0075
		$\tau = .02$	1.029	.0041
	Average n_i = 400	$\tau = 0$	1.379	.0017
		$\tau = .005$	1.016	.0007
		$\tau = .02$	1.002	.0000

MSE_{CC} to MSE_{MLE} Ratios

Table 5.67 summarizes the MSE ratios across all simulation conditions. As was the case with the MCAR data with continuous predictors, on average maximum-likelihood estimation provides large gains in efficiency, especially for the estimation of τ. As shown in Table 5.68, the actual gain in efficiency varies depending on the values of the simulation hyperparameters.

131

Table 5.67

	β_0	β_1	β_2	β_3	τ
Mean	1.621	1.622	1.269	1.312	2.281
Median	1.468	1.515	1.243	1.245	1.818
Minimum	1.159	1.192	1.063	1.083	1.066
Maximum	2.160	2.152	1.654	1.682	5.682

MSE_{CC}/MSE_{MLE} Ratios (MCAR Data w/Dichotomous Predictor) (header spanning above)

Table 5.68

Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.629	1.637	1.367	1.410	3.432
	.005	1.636	1.638	1.263	1.316	1.959
	.02	1.599	1.591	1.176	1.211	1.451
Avg. n_i	80	1.686	1.669	1.289	1.312	2.591
	400	1.556	1.575	1.248	1.312	1.971
Predictor inter-corrs.	Low	1.587	1.619	1.311	1.348	2.171
	High	1.655	1.625	1.226	1.276	2.390
V_{mod}	.006	1.688	1.684	1.268	1.291	2.461
	.03	1.554	1.560	1.269	1.334	2.100
Incidence of M. Data	50%	1.361	1.351	1.176	1.199	1.861
	75%	1.882	1.893	1.361	1.425	2.700

MSE_{CC}/MSE_{MLE} Ratios, Main Effects (MCAR Data w/Dichotomous Predictor) (header spanning above)

As with the data with all continuous predictors, the superiority of MLE estimation to CC estimation is greater when there is more missing data. The improvement is greatest for estimation of β_0, β_1, and τ. The relative efficiency of the EM estimates of β_2, β_3, and τ is lower for large values of τ; this finding mirrors that found in the previous sections as well. Unlike in the previous sections, there were no substantively large interactions between simulation hyperparameters regarding the size of the relative efficiencies of the $\hat{\beta}$s. There were three substantively large interaction for the relative efficiency for $\hat{\tau}$.

Table 5.69 demonstrates that the interaction between the size of τ and the incidence of missing data is such that when there is 75% missing data, the ML method is especially efficient for low values of τ, especially $\tau = 0$.

Table 5.70 shows that for a small average study sample size, the ML method is especially efficient for $\tau = 0$ and $\tau = .005$. For a larger average study sample size, the ML method is only especially efficient (i.e., three to four times as efficient) for $\tau = 0$.

Table 5.71 shows that for a small average study sample size, the ML method is especially efficient for 75% missing data.
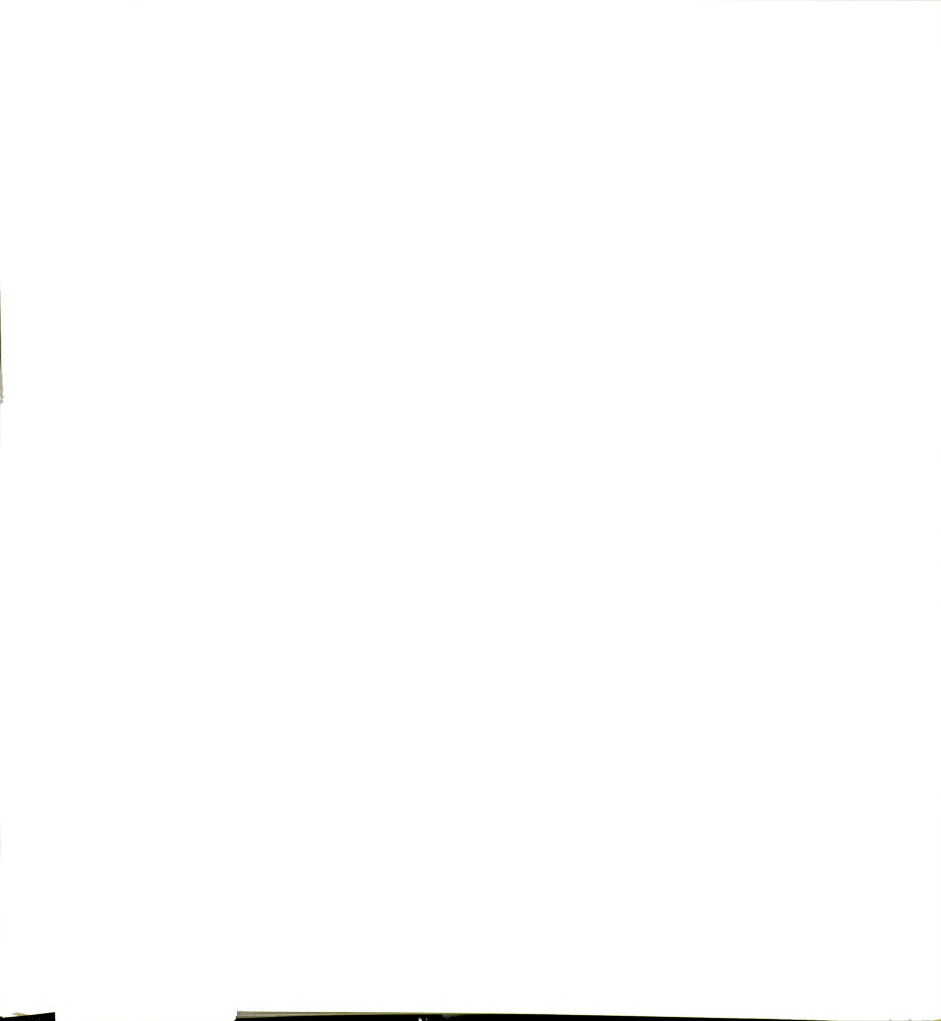
Table 5.69

MSE_{CC}/MSE_{MLE} Ratios for τ, MCAR Data w/Dichotomous Predictor		
(Size of τ x Incidence of Missing Data)		
		Ratio for τ
50% Incidence of Missing Data	$\tau = 0$	2.684
	$\tau = .005$	1.590
	$\tau = .02$	1.309
75% Incidence of Missing Data	$\tau = 0$	4.179
	$\tau = .005$	2.328
	$\tau = .02$	1.593

Table 5.70

MSE_{CC}/MSE_{MLE} Ratios for τ, MCAR Data w/Dichotomous Predictor		
(Size of τ x Average Study Sample Size)		
		Ratio for τ
Average $n_i = 80$	$\tau = 0$	3.678
	$\tau = .005$	2.635
	$\tau = .02$	1.460
Average $n_i = 400$	$\tau = 0$	3.186
	$\tau = .005$	1.484
	$\tau = .02$	1.443

Table 5.71

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR Data w/Dichotomous Predictor (Incidence of Missing Data x Number of Studies)		
		Ratio for τ
50% Incidence of Missing Data	Average n_i = 80	1.989
	Average n_i = 400	1.734
75% Incidence of Missing Data	Average n_i = 80	3.193
	Average n_i = 400	2.207

MSE$_{CC}$ to MSE$_{AC}$ Ratios

Table 5.72 summarizes the MSE ratios across all simulation conditions.

Table 5.72

MSE$_{CC}$/MSE$_{AC}$ Ratios (MCAR/MAR Data)					
	β_0	β_1	β_2	β_3	τ
Mean	1.333	1.325	1.001	.997	1.044
Median	1.254	1.261	1.003	.990	.967
Minimum	.791	.804	.673	.704	.456
Maximum	1.796	1.933	1.349	1.302	1.666

As with the MCAR data with all continuous predictors, available-case methods provide better estimates than complete-case methods for β_0 and β_1 and approximately

equally efficient estimates for β_0, β_1, and τ. Table 5.73 summarizes the MSE ratios by main effect.

Table 5.73

MSE$_{CC}$/MSE$_{AC}$ Ratios, Main Effects (MCAR Data w/Dichotomous Predictor)						
Parameter	Parameter Value	β_0	β_1	β_2	β_3	τ
τ	0	1.200	1.188	.895	.930	1.053
	.005	1.338	1.326	.998	1.018	1.142
	.02	1.460	1.462	1.100	1.183	1.270
Avg. n_i	80	1.429	1.409	1.053	1.036	1.209
	400	1.236	1.241	.949	.959	.879
Predictor inter-corrs.	Low	1.327	1.368	1.070	1.067	1.018
	High	1.338	1.282	.931	.928	1.069
V_{mod}	.006	1.377	1.375	1.004	1.006	1.084
	.03	1.288	1.275	.998	.989	1.003
Incidence of M. Data	50%	1.117	1.104	.930	.924	.819
	75%	1.547	1.546	1.071	1.071	1.268

As was the case with the MCAR data with all continuous predictors, AC estimation tends to be more efficient relative to CC estimation when combinations of hyperparameters are such that they make correlations among predictors and between

136

predictors and outcome lower. Correlations are lower for high τ, low average n_i, low correlations among predictors, and low values of V_{mod}; relative efficiencies are higher (though sometimes only marginally so) for these combinations of hyperparameters. The largest effect on relative efficiency is from incidence of missing data; the efficiency of AC estimation is almost indistinguishable from CC estimation when there is less missing data.

Table 5.74

MSE$_{CC}$/MSE$_{MLE}$ Ratios for τ, MCAR data w/Dichotomous Predictor (Average Study Sample Size x Size of τ)		
		τ Ratio
Average n_i = 80	$\tau = 0$	1.204
	$\tau = .005$	1.202
	$\tau = .02$	1.219
Average n_i = 400	$\tau = 0$	.655
	$\tau = .005$	.834
	$\tau = .02$	1.147

There is one substantively important interaction effect for the relative efficiency in the estimation of τ. Table 5.74 demonstrates that for large average study sample size, the relative efficiency is strongly dependent on the value of τ, while the efficiency is not dependent in meta-analyses with a small average study sample size.

The evidence in Tables 5.68 and 5.73 show that the efficiency of ML estimation relative to CC estimation decreases as τ increases, while the efficiency of AC estimation

relative to CC estimation increases as τ increases. Table 5.75 shows the average relative efficiency of the ML estimation to AC estimation for $\tau = .02$. These results are also similar to those for the MCAR data with continuous predictors. Maximum-likelihood estimation of τ is on average 24% more efficient for these conditions, but the estimation of the βs is on average only 7 to 10 percent more efficient.

Table 5.75

MSE_{AC}/MSE_{MLE} Ratios, $\tau = .02$ (MCAR Data, w/Dichotomous Predictor)					
	β_0	β_1	β_2	β_3	τ
Mean	1.096	1.093	1.071	1.105	1.244
Median	1.095	1.101	1.069	1.110	1.217
Minimum	1.056	1.021	.943	.971	1.046
Maximum	1.135	1.133	1.179	1.207	1.463

There is one condition out of the 48 for which ML estimation of β_2 and β_3 is marginally less efficient than AC estimation. In this condition, $\tau = .02$, average $n_i = 40$, $V_{mod} = .006$, and there are correlations of .10 among predictors. This is the condition for which the correlation between the predictors and the outcome is the lowest; the sample R^2 for this condition averaged .075. The relative efficiency of ML to AC estimation of τ was 1.17 for this set of hyperparameters. The other 47 combinations of hyperparameters had MSE ratios of above 1.00 for each hyperparameter.

6. Results: Bootstrapped Standard Errors

Bootstrapping Errors for the Slopes

Tables 5.76 and 5.77 show the main effects of the simulation parameters on the size of the bootstrapped variance/MSE ratios. Ratios based on both the mean bootstrapped variance across 100 simulations and median bootstrapped variance across the 100 simulations are reported due to the large effect that outliers had on the mean for some conditions in which $k = 40$ and the incidence of missing data was 75%. While most bootstrapped variances for slopes for those conditions were between .25 and 1.5, on some occasions the bootstrapped variance was over 5.00, and in one case it was over 50, even though the estimations still converged. Further investigation into the behavior of these estimates should lead to some rules of thumb for bootstrapping that will exclude unreasonably large values; for the purpose of this investigation, concentration will be focused on the performance of the ratios based on the medians.

These results are similar to what is found in Su (1988). A direct comparison is difficult, given that Su used different missing data percentages and underlying covariance matrices. For his condition in which there are 40 cases (equivalent to $k=40$ in this simulation study), and a 62.5% incidence of missing data (i.e., only 15 of the studies had complete data), he found an average ratio of 1.77 across all slopes. As a comparison, the average ratio for the condition in this study where $k=40$ and there was a 75% incidence of missing data was 2.617 (Table 5.78). The average ratio was considerably lower – only 1.170 – when there was a 50% incidence of missing data. Thus, the average ratio for the $k=40$ conditions was 1.894, similar to Su's average ratio.

139

For his condition in which there are 160 cases, and a 62.5% incidence of missing data, the average ratio was 1.03, which is similar to what was found in this investigation for both the 75% incidence of missing data condition and 50% incidence of missing data condition, when $k=100$.

There was only one interesting interaction, between incidence of missing data and number of studies, and it is also reflected in Table 5.78. Ratios using the medians are very close to 1.00 for all conditions except $k = 40$ and 75% incidence of missing data.

Table 5.76

Average Mean and Median Bootstrapped Variance/MSE$_{ML}$ Ratios for β_0 and β_1, Main effects					
Parameter	Parameter Value	Ratio for β_0 (Mean)	Ratio for β_0 (Median)	Ratio for β_1 (Mean)	Ratio for β_1 (Median)
τ	0	1.503	1.161	1.434	1.100
	.005	1.389	1.134	1.430	1.147
	.02	1.273	1.076	1.345	1.097
k	40	1.753	1.249	1.802	1.274
	100	1.023	.998	1.010	.956
Predictor inter-corrs.	Low	1.344	1.089	1.406	1.126
	High	1.432	1.578	1.406	1.103
Avg. n_i	80	1.424	1.140	1.500	1.154
	400	1.353	1.107	1.313	1.076
V_{mod}	.006	1.438	1.149	1.422	1.119
	.03	1.339	1.098	1.391	1.111
Incidence of M. Data	50%	1.040	.984	1.076	.997
	75%	1.736	1.263	1.737	1.232
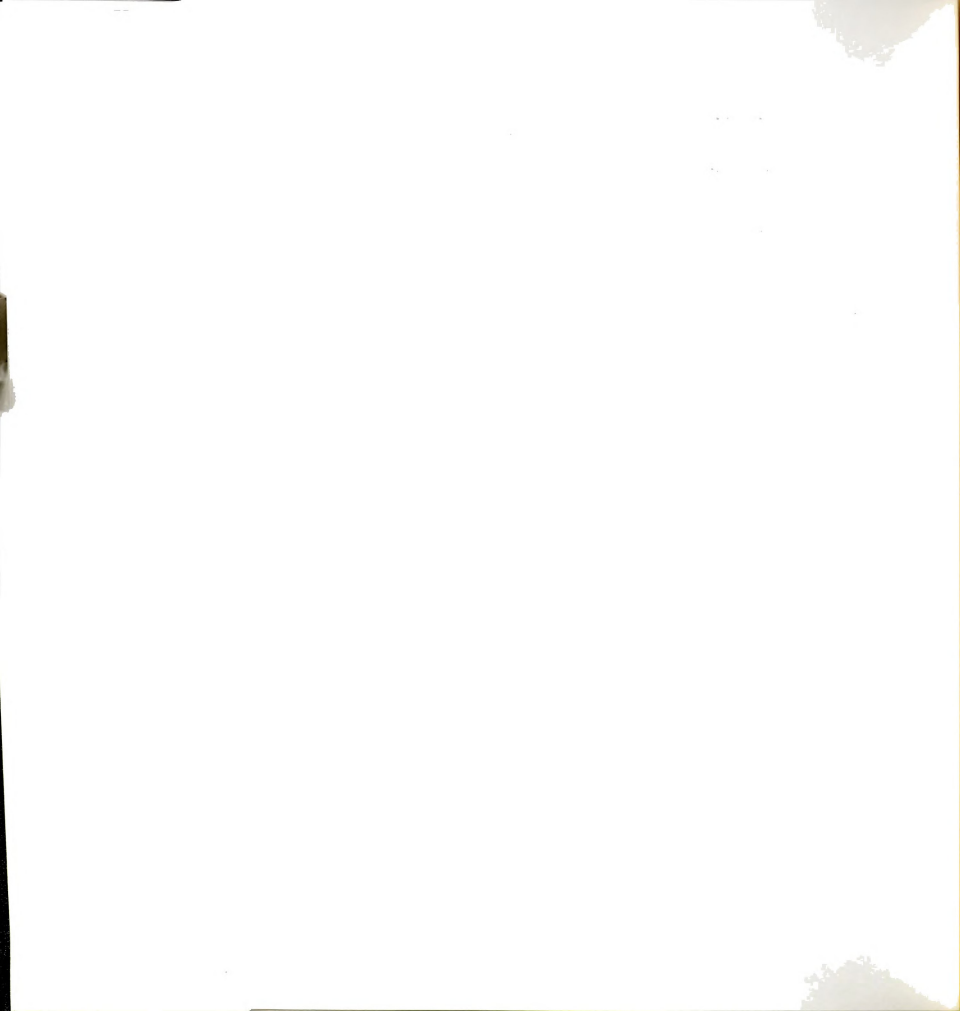
Table 5.77

Average Mean and Median Bootstrapped Variance/MSE$_{MI}$ Ratios for β_2 and β_3, Main effects					
Parameter	Parameter Value	Ratio for β_2 (Mean)	Ratio for β_2 (Median)	Ratio for β_3 (Mean)	Ratio for β_3 (Median)
τ	0	1.716	1.192	1.687	1.132
	.005	1.289	1.032	1.368	1.075
	.02	1.370	1.033	1.447	1.022
k	40	1.871	1.223	2.007	1.254
	100	1.045	.9481	.994	.900
Avg. n_i	80	1.500	1.091	1.570	1.095
	400	1.416	1.164	1.431	1.059
Predictor inter-corrs.	Low	1.546	1.128	1.590	1.078
	High	1.371	1.044	1.410	1.076
V_{mod}	.01	1.472	1.117	1.456	1.077
	.03	1.444	1.054	1.545	1.077
Incidence of M. Data	50%	1.090	.995	1.913	.971
	75%	1.827	1.177	1.088	1.183

Table 5.78

Average Mean and Median Bootstrapped Variance/MSE$_{ML}$ Ratios
(Incidence of Missing Data x Number of Studies Interaction)

		Ratio for β_0 (Mean)	Ratio for β_0 (Median)	Ratio for β_1 (Mean)	Ratio for β_1 (Median)	Ratio for β_2 (Mean)	Ratio for β_2 (Median)	Ratio for β_3 (Mean)	Ratio for β_3 (Median)
50% Incidence of Missing Data	$k = 40$	1.135	1.034	1.155	1.042	1.147	1.015	1.207	1.039
	$k = 100$	.945	.926	.996	.953	1.033	.975	.968	.902
75% Incidence of Missing Data	$k = 40$	2.371	1.455	2.449	1.506	2.596	1.432	2.806	1.469
	$k = 100$	1.102	1.072	1.025	.958	1.058	.921	1.020	.898

Investigations of the average non-coverage rates lead to similar conclusions. Tables 5.79 and 5.80 show, respectively, the main effects and incidence x k interaction effect on the frequency of non-coverage. A 5% rate is expected if the standard errors are being correctly estimated; the "non-coverage rate" nomenclature is used by Su (1988), but it also might be thought of as a rejection rate. Most effects are quite small. It appears that for smaller amounts of missing data, the non-coverage rate is slightly conservative; when there are larger amounts of missing data, the non-coverage rate may be slightly conservative for large k but slightly liberal for small k.

Overall, it should be concluded that bootstrapped standard errors for the slopes lead to rejection rates close to .05 and standard errors that are generally the correct size, though they may be a bit too large for meta-analyses with large amounts of missing data and few studies.

Table 5.79

Non-Coverage Rates using Bootstrapped Standard Errors, Main effects					
Parameter	Parameter Value	N.C. Rate for β_0	N.C. Rate for β_1	N.C. Rate for β_2	N.C. Rate for β_3
τ	0	5.9%**	5.3%	5.2%	5.0%
	.005	5.6%	5.4%	6.5%**	5.6%
	.02	6.2%**	5.9%**	6.6%**	6.5%**
k	40	5.3%	4.7%	5.3%	4.8%
	100	6.5%**	6.4%**	6.9%**	6.6%**
Avg. n_i	80	5.7%	5.7%	6.7%**	5.6%
	400	6.0%**	5.4%	5.5%	5.8%**
Predictor Inter-Corrs.	Low	5.6%	5.4%	5.6%	5.7%
	High	6.2%**	5.7%	6.6%**	5.7%
V_{mod}	.01	5.5%	5.3%	6.2%**	5.9%**
	.03	6.2%**	5.8%	6.0%	5.5%
Incidence of Missing Data	50%	6.7%**	6.2%**	6.7%**	6.4%**
	75%	5.0%	4.9%	5.4%	5.0%

** Non-coverage rate is significantly different from .05
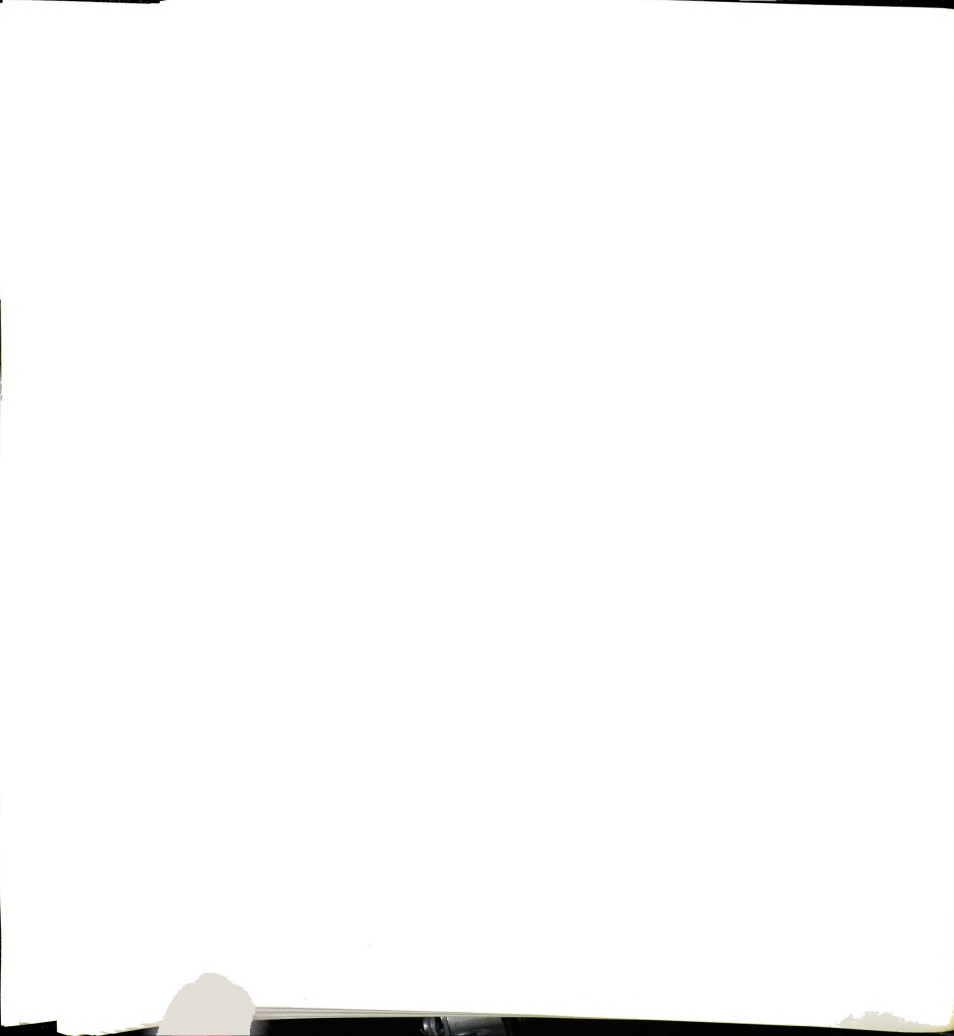
Table 5.80

Non-Coverage Rates using Bootstrapped Standard Errors					
(Incidence of Missing Data x Number of Studies Interaction)					
		N.C. Rate for β_0	N.C. Rate for β_1	N.C. Rate for β_2	N.C. Rate for β_3
50% Incidence	$k = 40$	6.5%**	5.8%	7.2%**	5.9%**
of Missing Data	$k = 100$	6.9%**	6.8%**	6.8%**	6.9%**
75% Incidence	$k = 40$	4.1%	3.6%**	3.4%**	3.8%**
of Missing Data	$k = 100$	6.0%**	6.1%**	7.0%**	6.3%**

** Non-coverage rate is significantly different from .05

<u>Testing Homogeneity of Effects</u>

The test of $H_0 : \tau = 0$ is an important one in any meta-analysis. Unfortunately, the negative bias in the estimation of τ together with the "floor" effect (disallowing estimates of $\tau < 0$) combine to make bootstrapped standard errors of the estimate of τ smaller than they should be, leading to many empirical rejection rates over 20%. A value of $\hat{\tau}$ was considered to be significantly above zero if it exceeded 1.675 times the corresponding bootstrapped standard error (i.e., the test was one-tailed, as τ cannot be less than zero.) The estimates of τ should be asymptotically normal, given the nature of maximum-likelihood estimation. The results are summarized in Table 5.81.

Table 5.81

Empirical Rejection Rates using Bootstrapped Standard Errors for Test of $H_0: \tau = 0$		
Parameter	Parameter Value	Rejection Rate
k	40	25.1%**
	100	23.0%**
Avg. n_i	80	22.4%**
	400	25.6%**
Predictor Inter-Corrs.	Low	20.0%**
	High	28.1%**
V_{mod}	.006	29.6%**
	.03	18.5%**
Incidence of Missing Data	50%	23.0%**
	75%	25.1%**

** Non-coverage rate is significantly different from .05

While the empirical rejection rates are clearly above the nominal ones, the problem is, practically speaking, a small one, given the phenomenal precision of the ML estimator of τ when $\tau = 0$. Average medians, 90th, and 95th percentiles for the 32 combinations of hyperparameters for which the data were MCAR and the average study sample size was

80 are presented in Table 5.82. The value of τ was estimated extraordinarily well for the 32 combinations for which the sample size was 400; the largest 95th percentile across all of these conditions was .0028.

Table 5.82

Average Empirical Percentiles for Estimates of τ, Average n_i=80, $\tau = 0$					
Parameter	Parameter Value	Median	75th Percentile	90th Percentile	95th Percentile
k	40	.0001	.0014	.0042	.0088
	100	.0001	.0020	.0054	.0082
Predictor inter-corrs.	Low	.0001	.0016	.0046	.0089
	High	.0001	.0018	.0053	.0078
V_{mod}	.006	.0001	.0015	.0043	.0075
	.03	.0001	.0020	.0055	.0094
Incidence of M. Data	50%	.0001	.0018	.0057	.0091
	75%	.0001	.0016	.0042	.0069

Across all conditions, when $\tau = 0$, the typical estimate of τ is generally very close to zero, even when the average study sample sizes are low. Improvement in the statistical significance test of the hypothesis H_0: $\tau=0$ will be pursued in future research.

CHAPTER VI

SAMPLE META-ANALYSIS

To demonstrate that the program written to implement the EM estimation procedure derived in Chapter III could handle real data, a meta-analysis was conducted on data supplied by Professor Mark Lipsey of Vanderbilt University. The dataset contained 3905 effect sizes representing the efficacy of interventions designed to reduce the extent of juvenile delinquency among at-risk youth or youth who had already committed some delinquent acts. Over 150 study characteristics were coded for each effect. The most recent analyses of this dataset can be found in Lipsey (1999a) and Lipsey (1999b). Due to the complexity of the data, problems that specifically relate to the nature of the problem involved (i.e., the types of study characteristics investigated in juvenile delinquency studies), and more importantly, less familiarity on my part with the both the dataset and the subject matter than Dr. Lipsey, a full replication of his analyses could not be conducted. However, a careful reading of Lipsey (1999a, 1999b) suggests a model that might be examined in a test of the EM program.

This chapter begins with a description of how the sample of study effects was selected from the database, and continues with a description of the study characteristics to be investigated in this mixed-model meta-analysis. Listwise-deletion (complete-case) parameter estimates and EM parameter estimates for the initial model and the final model are given and the differences between the results of the two models are discussed.

1. Selection of Study Effects and Study Characteristics

While the Lipsey database included 3905 effect sizes, many effect sizes were dependent on one another. Most studies had both post-test and follow-up measurements and multiple outcomes. The most straightforward way to eliminate these dependencies was to pick one effect size from each study. While not all studies had follow-up measurements, all studies did have post-test measures; thus, only effect sizes based on post-test measures were used. Lipsey (1999b) limits his analysis to studies that have a recidivism outcome measure; preferably police arrests, but if that was not available as an outcome the most similar outcome was used. In this meta-analysis police arrests was used as the preferred outcome; when that was unavailable, "institutionalization" was used, and when that was not available, probation, court, or parole contact was used. Lipsey also limits his analysis to studies in which the researcher did not implement the treatment, finding that such studies tended to be biased; the same restriction was made here. Finally, studies that did not have complete data on the sample size of the treatment and control groups were also eliminated. This procedure led to a dataset consisting of 328 studies.

Lipsey found many variables important in his 1999 analyses. However, not all of the variables that Lipsey found to be important in his sample had even small bivariate relationships with the effect sizes in this sample; this is another reason that this analysis cannot be considered to be a replication of his analysis. For example, Lipsey found that four types of treatments related to probation led to higher treatment effect sizes; the same was not found in the sample used in this meta-analysis. Nevertheless, most of the variables Lipsey found to be important were considered in the initial model below.

Table 6.1

Study Characteristics Investigated in Juvenile Delinquency Meta-Analysis

Mean Age of Juveniles in Program at Time of Intervention
 • Ranged from 10.9 to 21.0

Aggressive History of Juveniles
 • 1 = At least some aggressive history, 0 = No aggressive history

Administrator of the Treatment
 • 1 = Administrator was criminal justice personnel, 0 = Administrator was not criminal justice personnel

Reason for Entering the Treatment
 • 1 = Admission was mandatory, 0 = Admission was not mandatory

Site of the Treatment
 • 1 = Site was a criminal justice site, 0 = Site was not a criminal justice site

Periodicity of the Treatment
 • 1 = Treatment took place either Daily or more infrequently, 0 = Treatment was "continuous" in nature

Amount of Weekly Contact
 • 1 = Average number of weekly hours > 7.0, 0 = Average number of weekly hours < 6.99

Length of Program
 • 1 = Program lasted 18 weeks or more, 0 = Program lasted 17.9 weeks or less

Deterrence/Wilderness Treatment
 • 1 = Program was either based on deterrence or a wilderness camp or survival training, 0 = All other programs

Counseling Treatment
 • 1 = Program's central focus was some form of individual or group counseling, 0 = All others

Demonstration Program
- 1 = Program was a "Demonstration program", 0 = All others (e.g., public or private programs)

Private Program
- 1 = Program was privately sponsored, 0 = All others

Difficulty in implementation of program
- 1 = Program was difficult to implement, 2 = Program was possibly difficult to implement, 3 = Program was not difficult to implement

Table 6.2

Variable Names and Frequency of Missing Data	
Variable	Proportion of Missing Data
MeanAge	38/328 (11.6%)
AggrHist	105/328 (32.0%)
Adminstr	0/328 (0%)
Mandatory	4/328 (1.2%)
CJSite	22/328 (6.7%)
Periodicity	51/328 (15.5%)
WeeklyContct	22/328 (6.7%)
ProgLength	46/328 (14.0%)
Springer	0/328 (0%)
Counsel	0/328 (0%)
DemoProg	2/328 (.6%)
PrivProg	2/328 (.6%)
Difficulty	23.328 (7.0%)

Table 6.1 describes the thirteen variables used in the first model tested; all variables except for "history of aggression" and "difficulty in implementing treatment" were analyzed by Lipsey.

The variable names in Table 6.2 correspond, in the same order, to the variables described in Table 6.1. (The Deterrence/Wilderness Camp/Survival Camp variable is named *Springer* due to the frequent appearance of advocates of these kinds of programs on the Jerry Springer TV show.) Note that all variables investigated except for *MeanAge*

and *Difficulty* are dichotomous in nature. Two characteristics that at first glance might best be considered continuous – program length and mean hours of weekly contact – were dichotomized by Lipsey (and by this researcher) due to the fact that the distributions of values for those variables were highly skewed. Most programs lasted fewer than 20 weeks, but some programs lasted, according to the database, over 5 years. Similarly, while the majority of values for average hours of weekly contact was under 8 hours, many interventions took over 40 hours in a week. While section 7 of Chapter V showed positive results for models containing a single dichotomous predictor that had missing values, and most of these predictors had either no or low amounts of missing data (and it is generally accepted that dichotomous variables with no missing data are acceptable in EM estimation), it should be kept in mind that the EM estimation derived in Chapter III is based on normally distributed predictors.

2. The Initial Model

The first model investigated had as its outcome the effect size and as its predictors the values of the 13 study characteristics mentioned above. The complete-case analysis used 146 studies, while the EM analysis used all 328. This shows the usefulness of missing-data methods even when the dataset is not very sparse, as is the case with this collection of studies. Even though only one variable has a missing data percentage above 20%, and the majority of variables have missing data percentages less than 15%, over half the studies were lost when all cases with *any* missing data were dropped. The mean ES was calculated by adding the estimate of β_0 to the sum of the products of the estimates of
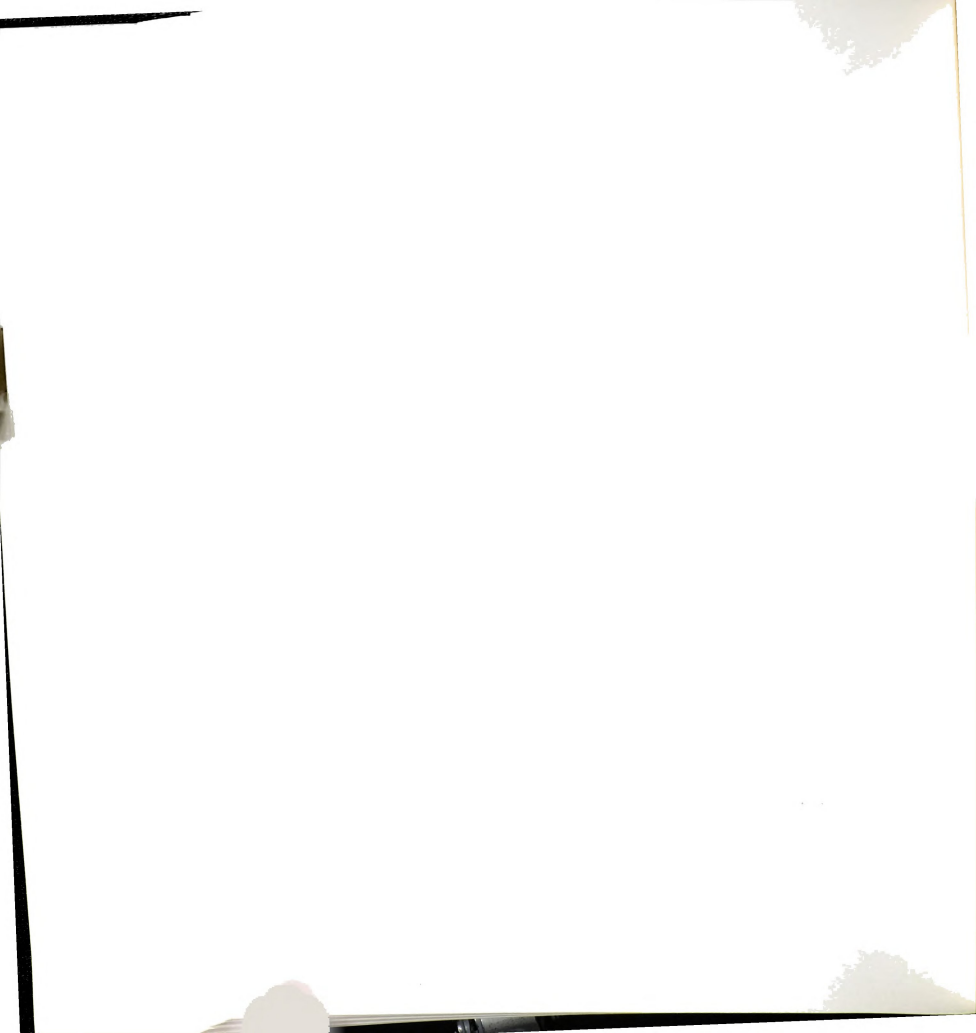
the slopes of the predictors and the weighted mean value of those predictors.

Table 6.3

Parameter Estimates and Significance Tests: Initial Model				
Parameter	CC Estimate w/standard error	CC p-value	EM estimate w/standard error	EM p-value
Mean ES	.1428 (.0258)	<.001	.1470 (.0175)	< .001
MeanAge	-.0271 (.0163)	.096	-.0242 (.0108)	.024
AggrHist	-.0464 (.0641)	.470	-.0421 (.0484)	.383
Adminstr	-.0550 (.0810)	.498	-.0281 (.0407)	.491
Mandatory	.1470 (.0639)	.022	.0606 (.0391)	.121
CJSite	-.0737 (.0797)	.355	-.0643 (.0314)	.041
ProgLength	.0694 (.0552)	.209	.0125 (.0367)	.733
Periodicity	.0794 (.0729)	.276	.0513 (.0530)	.333
WeeklyCont	-.0473 (.0693)	.495	-.0395 (.0509)	.437
Springer	-.0890 (.1223)	.467	-.2434 (.0638)	<.001
Counsel	.0230 (.0807)	.776	.0476 (.0539)	.377
DemoProg	-.0040 (.0677)	.953	.0141 (.0392)	.718
PrivProg	.0386 (.1011)	.703	.0925 (.0522)	.076
Difficulty	.0476 (.0320)	.137	.0349 (.0211)	.098
τ	.0584	<.001	.0573	<.001[4]

---

[4] There is no strict test of H_0: $\tau = 0$ using the EM method, but as shown in Chapter V, the EM estimate of τ is usually far more accurate than the CC estimate.

155

Table 6.3 shows the parameter estimates and significance test results for both the complete-case and EM estimations. The CC estimations find that the mean effect size is significantly different from zero, although there is a large amount of random-effects variation. Only one other parameter is found to be significant – *Mandatory*, with a parameter value indicating a difference in average effect size of .147 favoring programs in which participation was mandatory as opposed to voluntary. Two other variables, *MeanAge* and *Difficulty* tend towards marginal significance (*p*-values of .097 and .137, respectively).

The EM estimation tells a slightly different story. As with the CC estimation, the mean effect size is significantly different from zero, and there is a large amount of random-effects variation. However, the EM estimates find three variables to have slopes statistically significantly different from zero: *MeanAge*, *CJSite*, and *Springer*. The difference stems from the lower standard errors from the EM method; if the CC estimates remained the same, but had the lower standard errors of the EM method, the CC estimates for *MeanAge* and *CJsite* would be statistically significant as well. The EM slopes for *PrivProg*, *Difficulty*, and *Mandatory* are all marginally statistically significant (with *p*-values of .076, .098, and .121, respectively). If the CC estimates remained the same but had the EM standard errors, the estimate for the slope for *Difficulty* would be statistically significant as well. Overall, the standard errors for the CC estimates range from 133% (for *AggHist*) to 254% (for *CJsite*) the size of the corresponding EM standard errors. Of course, these calculations make the assumption that the bootstrapped EM standard errors are accurate, but the findings for $k = 100$ with 50% missing data in section 8 of Chapter V

156
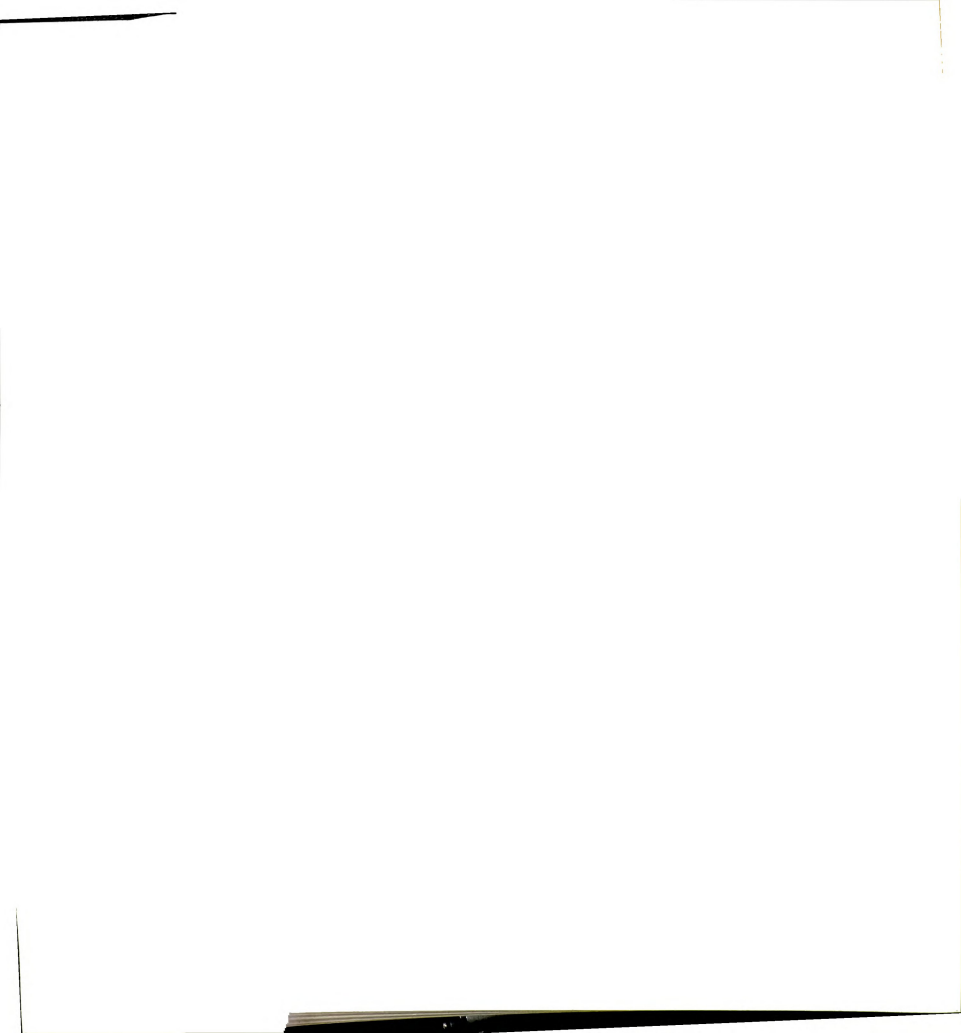
suggest that the standard errors for this condition are generally the correct size and on average lead to close to the correct confidence-interval coverage percentages.

3. The Final Model

The final model was determined by the less-than-theoretically-defensible method of simply dropping insignificant variables (from the EM estimation) one by one until all variables remaining were either statistically significant or at least marginally statistically significant. Each variable dropped was then tested again by adding it to the final model, alone, to see whether its slope became statistically significant or marginally statistically significant. Obviously, a more careful, theory-backed method of model testing should be used in a future application of EM estimation to this dataset, but it seemed unwarranted at this juncture given the purpose of this analysis (which was not to replicate a specific Lipsey analysis), the fact that the sample of 328 effects was clearly not the sample of effects used in either Lipsey (1999a) or Lipsey (1999b), and the fact that such an analysis should be conducted in partnership with a subject-matter expert.

The final model had four predictors: *MeanAge*, *Periodic*, *Springer*, and *PrivSpon*. CC and EM parameter estimates are presented in Table 6.4.

Table 6.4

Parameter Estimates and Significance Tests: Final Model				
Parameter	CC Estimate w/standard error	CC p-value	EM estimate w/standard error	EM p-value
Mean Effect Size	.1341 (.0188)	<.001	.1443 (.0149)	< .001
MeanAge	-.0315 (.0106)	.004	-.0340 (.0108)	.002
Periodicity	.1023 (.0391)	.006	.0945 (.0376)	.012
Springer	-.2327 (.0827)	.010	-.2768 (.0846)	.001
PrivProg	.1148 (.0649)	.130	.1218 (.0451)	.007
τ	.0522	<.001	.0590	τ<.001

The results indicate a very large negative effect for the *Springer* programs, similar to what Lipsey (1999a) found in his analysis. While the shock-scared and wilderness programs seem to be popular with the public, they work less well than other types of programs, and might actually have a detrimental effect in some instances. The result for *MeanAge* was also strong. For instance, holding other variables constant, treatment for juveniles of 12 years of age would on average have an effect size .27 better than treatment for juveniles of 20 years of age. While this result makes substantive sense (it makes sense that it would be easier to change behavior among the young than among the old), this result is the opposite of that found in Lipsey (1999a), where he found that treatments were more effective when the average age was greater than the median (15.5 years of age) for

158

all juveniles. The result in this analysis does not stem from the partial nature of the relationship; the unweighted bivariate correlation between *MeanAge* and the effect size is -.21. However, as noted above, while the sample used in this investigation is similar to that used in the Lipsey analyses, it is not identical. The effects for *PrivProg* and *Periodicity* are smaller, but substantively important given the overall small average effect size that juvenile delinquency treatments have. These effects are in line with those found in Lipsey's analyses; private programs tend to have larger effects than other programs, as do programs that deliver treatment in less than a continuous fashion.

The above findings are based on the EM estimates; however, the CC and EM estimates and significance tests are quite similar. This is not surprising, given that dropping the eight other predictors raised the number of complete cases considerably, from 146 to 256, leading to fully 78% of the studies having complete data. However, the improvement in standard error for *PrivProg* does lead to a statistically significant EM estimate of that slope, while the concomitant CC estimate is only marginally significant.

While the EM standard error for *PrivProg* is clearly superior to the CC standard error, the other EM standard errors for the other predictors are very similar to the CC standard errors for those same predictors. In fact, for two of the predictors (*MeanAge* and *Springer*), the EM standard errors are actually marginally larger than the CC standard errors.

4. Conclusions

The purpose of the simulation study was to test the estimation equations derived in Chapter 3; the purpose of this real data meta-analysis was to show that the program used to employ those equations could work in a non-simulation environment, with as many predictors as a meta-analyst might find pertinent to examine, and estimate slopes, standard errors for those slopes, and τ for that data using both complete-case and EM maximum-likelihood methods. In this respect, the purpose has been achieved, and in testing the program it was demonstrated that EM estimation can lead to important improvements in standard error in a real-data context in a model with many predictors. In the examination of the initial model's thirteen predictors, fully 55% of the data was incomplete, even though the missing data rates for most of the predictors were quite low. In this condition, assuming one trusts the EM bootstrapped standard errors (and more investigation needs to be done on the standard errors generated when most predictors are dichotomous), EM estimation of the parameters was at least 33% more efficient than CC estimation, and sometimes far more efficient. While the differences between the CC and EM estimations in the final model are far smaller, as would be expected given the large number of studies with complete data, the fact remains that EM estimation allows meta-analysts to include more sparsely measured variables (such as *AggrHist* here) in early models and have greater power to find significant results for all predictors.

Finally, it should be noted that the EM program used to estimate these models did not have to be customized in any but the most trivial way to handle the Lipsey dataset. Any ASCII meta-analytic dataset that obeys certain rules with respect to variable order

(such as what columns the study sample sizes or variance estimates are placed in) can be analyzed in short order. The program used to estimate the initial model can be changed to the program to estimate the final model in a matter of seconds, and the program can be changed to estimate models for a different dataset in a matter of minutes. The program outputs both CC and EM estimates, standard errors, 95% confidence intervals for each slope, and p-values for the tests of each slope, labeled by variable name. The user has the option of what variables to include in the analysis and how many bootstraps to use to find the EM standard errors. A copy of the program, which is written in SAS/IML, can be obtained by writing the author at fahrbach@msu.edu or contacting him through Dr. Betsy Becker at 456 Erickson Hall, Michigan State University, East Lansing, MI, 48824.
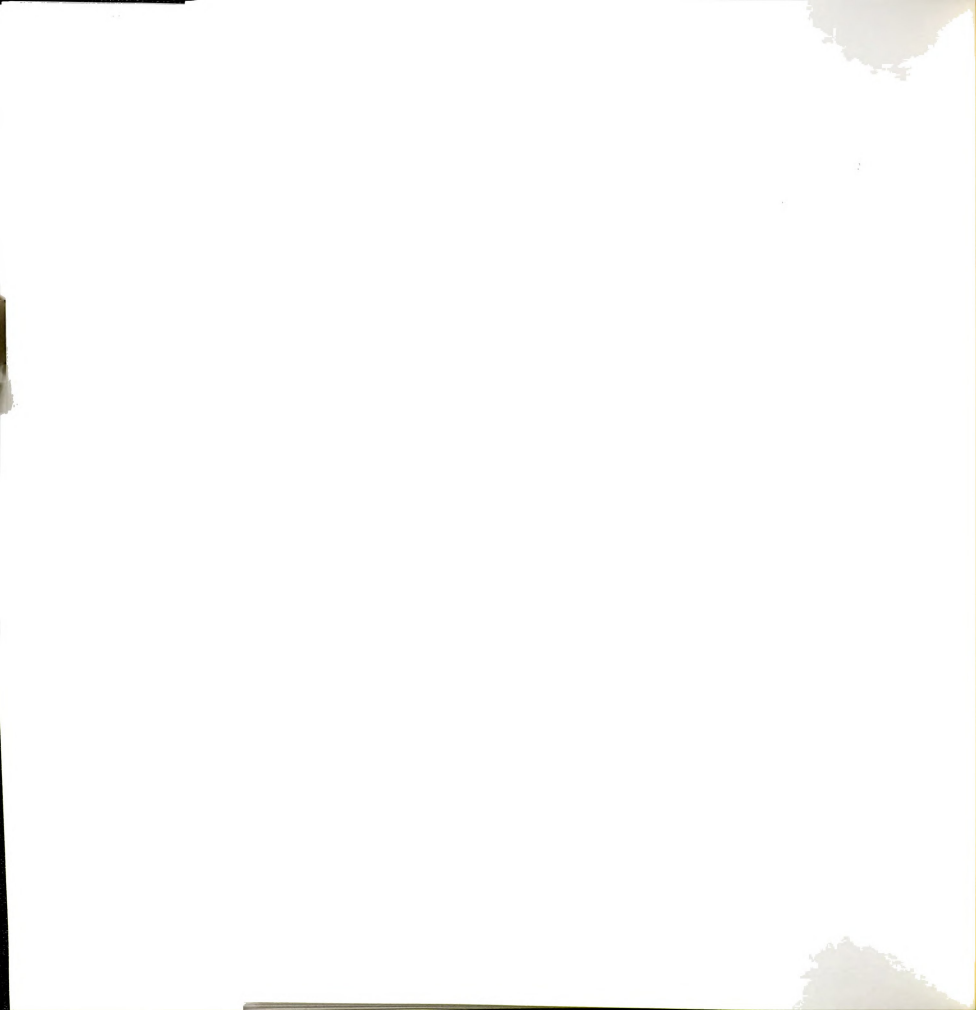
CHAPTER VII

DISCUSSION AND CONCLUSION

Maximum-likelihood estimation was expected to be more precise than available-case estimation and complete-case estimation. However, this does not necessarily imply that the maximum-likelihood method is always preferable, because it requires special software and is more difficult for non-statisticians to use and understand. The discussion in this chapter centers around two practical questions from the point of view of the non-statistician interested in conducting a meta-analysis:

1. If it is available and easy to use, should maximum-likelihood estimation always be used in meta-analysis? In other words, is maximum-likelihood estimation always better?

2. Are there situations in which CC or AC estimation is "good enough" because the ML method offers little additional efficiency in these situations? In other words, is maximum-likelihood estimation always *substantively* better?

1. Some Practical Considerations

Before addressing these questions, three pieces of information need to be considered. The first two consider the actual practice of meta-analysis (which differs from what transpired in the simulations in Chapter V), and the third, generalization of the results in Chapter V.

First, a meta-analysis is rarely as straightforward as one simple estimation of the
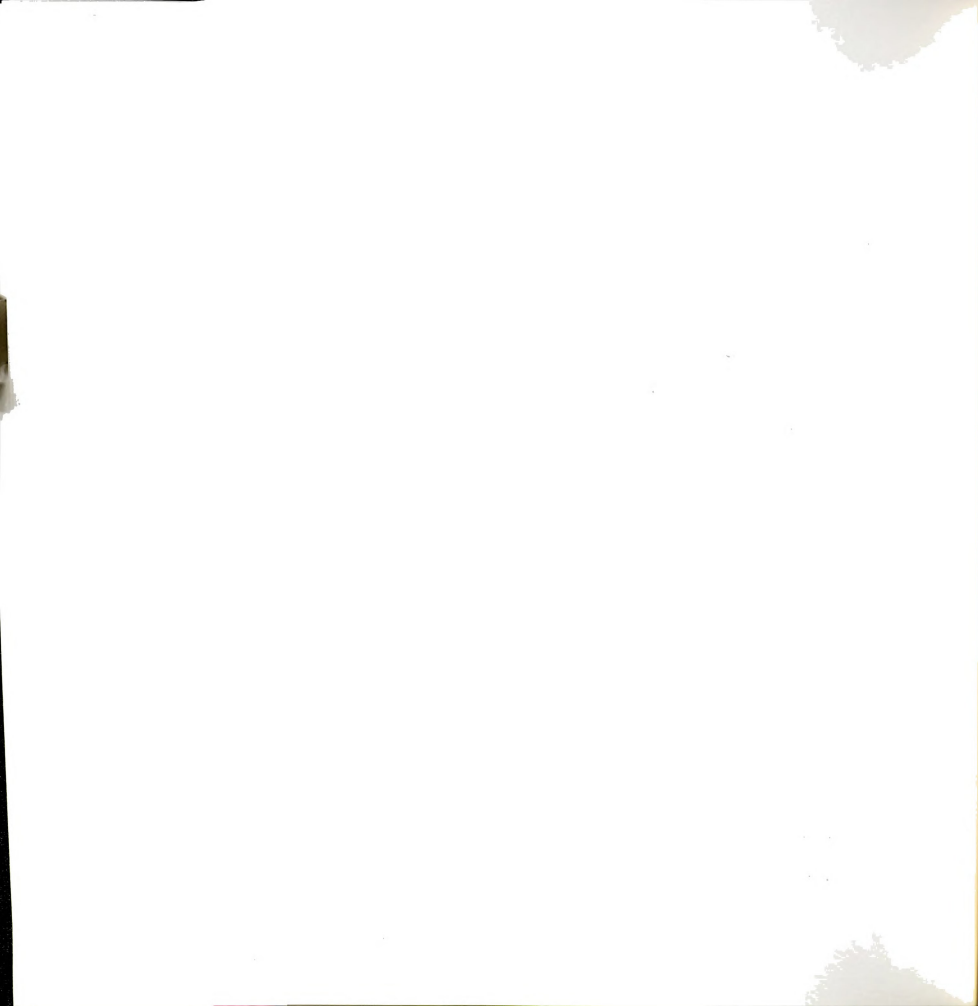
parameters across all studies of interest. In the analysis in Chapter VI, several models were estimated, as variables were dropped and (for the complete-case analysis) studies were added. The meta-analysis ended at that point, as its purpose was simply to show that the EM estimation program could handle real data. In a real meta-analysis, however, the analysis would only have begun. The final model could be tested within each level of each of the dichotomous variables in order to determine whether the effects were similar for any subsets of data. This is not an uncommon practice (e.g., see Lipsey, 1999b); it is done because of the possibility that for some subsets of studies, the predictors may explain most of the variation in the study effects, while in others the random-effects variation may dominate. Similarly, it is possible that there are interaction effects, and that predictors that are important within one subset of studies are not in another subset. Because the last thing a practitioner wants to do is use different estimation techniques for different subsets of data (e.g., use available-case estimation for analysis of the entire dataset and use complete-case or maximum-likelihood estimation for the subsets), I make the assumption that unless the group of studies in the meta-analysis is truly huge (perhaps over 1000) or unless the meta-analyst has good reason to exclude the possibility of analyses of subsets of data across levels of categorical variables, the meta-analyst is likely to desire methods that work for both small and large numbers of studies. Even if the number of studies is large, there is a strong possibility, as in the Lipsey dataset, that subsets of interest may have fewer than 50 studies.

The second consideration is that while seven simulation hyperparameters were studied, the meta-analyst is only going to know with certainty the values of three of them:

the number of studies, the average study sample size, and how much data is missing. The meta-analyst will only have sample values for the correlations between predictors and perhaps a rough estimate of the value of τ, both of which may change between subsets of data and may not accurately reflect the population values. The meta-analyst will have even less of an idea about what the values of V_{mod} might be in the data, and will likely have no idea whether the missing-data mechanism is such that the data are MCAR, NMAR, or NMAR (though many researchers will make the assumption that any missing data are probably MCAR in nature). Thus, when considering estimation techniques to recommend, I'll assume I know no more than a meta-analyst would know in this situation, i.e., the number of studies, the average study sample size, and whether missing data are (as defined in the simulations) sparse or heavy.

The third consideration is a caveat: while the statistics in Chapter V are informative, they may not generalize to all meta-analyses. While the simulations attempted to examine the behavior of CC, ML, and AC estimation across a wide range of conditions, there were restrictions. For instance, the simulations only looked at meta-analyses with three predictors, and only a partial analysis of the effects of categorical predictors was conducted. Thus, while it can be said that across the many conditions studied, ML estimation provided estimates of τ that were about three times as efficient as complete-case estimation, it obviously cannot be said that across the population of all meta-analyses, ML estimation of τ will be three times as efficient as CC estimation.

2. Is Maximum-Likelihood Estimation Always Better?

In the first question I temporarily neglect the complexity of use of maximum-likelihood estimation by non-statisticians by assuming that maximum-likelihood estimation for meta-analyses can be made easily accessible and employable. With this and the above three considerations in mind, question one is easy to address. In the simulations conducted, across all types of missing data (MCAR, MAR, and the three types of NMAR data), on average maximum-likelihood estimation provided (relative to complete-case estimation) roughly 100% more efficient estimation of β_0, 67% times more efficient estimation of β_1, 33% more efficient estimation of β_2 and β_3, and about 200% more efficient estimation of τ. The difference between the efficiencies for estimation for β_1 and the efficiencies for estimation of β_2 and β_3 seems to stem overwhelmingly from the fact that the first predictor was always observed, while the 2nd and 3rd predictors were only partially observed. The efficiencies for the 2nd and 3rd predictors were often very similar, even in the p-NMAR data, where only the 2nd predictor was used to generate the missing data pattern. In addition, large biases occurred in the complete-case estimation of the mean effect size when the data were MAR or NMAR, but bias was zero or near-zero for maximum-likelihood estimation. While substantial biases did not occur in the available-case estimation of the mean, for many combinations of hyperparameters (especially those with low numbers of studies and large degrees of missingness) available-case estimation was actually less efficient than complete-case estimation. While in some rare cases available-case estimation provided the most precise estimates of the three procedures for one or two parameters, even in these cases the difference in precision was marginal at

best, and maximum-likelihood estimation was more precise for the other parameters of interest. It cannot be said, then, that on average across the hyperparameters, available-case estimation was even marginally superior to maximum-likelihood estimation for any condition.

Because on average across the 528 conditions maximum-likelihood estimation provided as precise or (far more often) more precise estimates of the parameters than either of the competing methods, question one must be answered in the affirmative: maximum-likelihood estimation is always better, at least across the values of the hyperparameters studied.

3. Is Maximum-Likelihood Estimation Always Substantively Better?

For some, the answer to the first question is enough. Some meta-analytic databases require hundreds of thousands of dollars and much time to build and code (e.g., the Lipsey dataset), and researchers who have that much invested are likely to want to use the most advanced techniques possible to squeeze every last bit of precision into their estimates. Others, however, do not have as much invested, and may want more justification for moving from traditional complete-case methods to something more complex. These investigators may, justifiably, see little purpose in using a more complex method and specialized software unless it can be shown that the extra inconvenience incurred results in a good enough "payoff". While it is impossible to say precisely how much extra inconvenience ML estimation might be worth, or how much "payoff" a meta-analyst would look for, the only way to address this question is to somehow quantify the

166

improvement offered by EM methods.

"Payoff" has been chiefly measured in this paper as improvement in relative efficiency. Summaries of the average MLE to CC relative efficiencies and AC to CC relative efficiencies appear in Table 7.1 and Table 7.2. While the relative efficiencies were reported in Chapter V based on whatever simulation hyperparameters were the most important in the ANOVAS, here they are summarized for the eight conditions reflecting information available to a researcher, given his knowledge of the three hyperparameters mentioned above (number of studies, average study sample size, and incidence of missing data). The relative efficiencies given are those for MCAR data, as most meta-analysts assume that patterns of missingness in their studies are MCAR in nature. Only for MCAR data does complete-case estimation provide an unbiased estimator of the mean effect; if a meta-analyst had reason to suspect that his data was not MCAR, he would be *more* inclined to switch methods, not less.
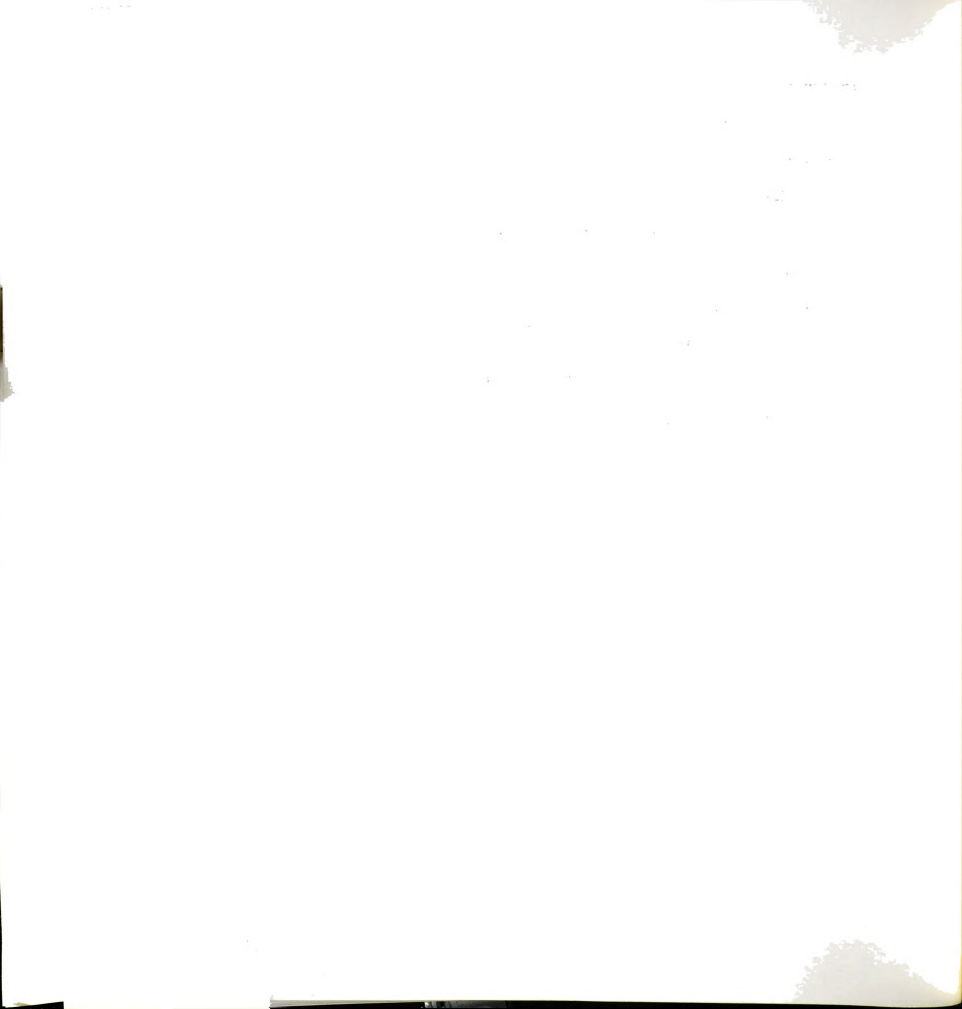
Table 7.1

MSE$_{CC}$/MSE$_{MLE}$ Ratios for MCAR Data

Across Hyperparameters Known to Meta-Analyst

			β_0	β_1	β_2	β_3	τ
50% Incidence of Missing Data	$k = 40$	Average $n_i = 80$	1.510	1.515	1.253	1.257	3.103
		Average $n_i = 400$	1.402	1.420	1.186	1.222	2.309
	$k = 100$	Average $n_i = 80$	1.392	1.375	1.174	1.180	1.953
		Average $n_i = 400$	1.335	1.323	1.168	1.190	1.760
75% Incidence of Missing Data	$k = 40$	Average $n_i = 80$	2.111	2.193	1.552	1.569	5.888
		Average $n_i = 400$	1.853	1.938	1.358	1.374	3.591
	$k = 100$	Average $n_i = 80$	2.008	1.987	1.411	1.434	3.156
		Average $n_i = 400$	1.803	1.778	1.322	1.415	2.327

Table 7.2

MSE_{CC}/MSE_{AC} Ratios for MCAR Data
Across Hyperparameters Known to Meta-Analyst

			β_0	β_1	β_2	β_3	τ
50% Incidence of Missing Data	$k = 40$	Average $n_i = 80$	1.303	1.247	1.003	.996	1.306
		Average $n_i = 400$	1.133	1.137	.934	.949	.874
	$k = 100$	Average $n_i = 80$	1.174	1.154	.959	.947	.927
		Average $n_i = 400$	1.036	1.038	.886	.892	.702
75% Incidence of Missing Data	$k = 40$	Average $n_i = 80$	1.255	.951	.621	.660	1.830
		Average $n_i = 400$	1.042	.718	.450	.435	1.110
	$k = 100$	Average $n_i = 80$	1.705	1.650	1.137	1.143	1.495
		Average $n_i = 400$	1.422	1.373	.977	1.000	1.075

The results in Tables 7.1 and 7.2 are as expected given the in-depth analysis in Chapter V. Maximum-likelihood estimation is always more efficient, often by a large margin. By contrast, there are often losses in efficiency for AC estimation, especially for β_2 and β_3. Also, AC estimation of τ is regularly worse than CC estimation of τ when there is a small amount of missing data. Before this study, it seemed possible that available-case estimation might provide a substantively acceptable way to avoid listwise deletion without having to resort to the complexity of maximum-likelihood methods. While, as noted above, available-case estimation seems to indeed provide precision similar to that of ML methods for some individual conditions, practically speaking, available-case estimation cannot be recommended.

Although available-case estimation cannot be recommended, it is not clear that the results in Table 7.1 lend unequivocal support for a ML meta-analysis, regardless of the added inconvenience (unless, as she should be, the meta-analyst is concerned with precise estimation of τ; the superiority of ML methods to estimate τ is transparent). For instance, consider the condition with 50% incidence of missing data, $k = 100$, and average $n_i = 400$. This might be considered the "worst case scenario" regarding improvement in efficiency over CC estimation, at least among the combinations of hyperparameters studied; the increases in efficiency for the estimation of β_2 and β_3 are less than 20%. This does not seem like much of an increase, and it is unclear what, substantively, such an increase in efficiency might mean.

This issue leads to another way to measure "payoff" – in terms of the power of the statistical tests. Both statisticians and non-statisticians alike are more used to power

concerns than concerns about relative efficiency, and with the mean-squared-errors available for CC and ML estimates of β it is straightforward to determine the power of each statistical test. Thus, for a given condition, the power of the CC and ML estimates of $\beta_0, \beta_1, \beta_2,$ and β_3 can be compared.

Specifically, for $\beta_1, \beta_2,$ and $\beta_3,$ the power of the test of H_O: $\beta = 0$ vs. H_A: $\beta = \beta_A$ was calculated (Mendenhall et al., 1986), where β_A represents the actual population value of the parameter in the simulations and the empirical root-mean-squared-error is used as the standard error of the estimate. For β_0, the power of the test H_O: $\beta_0 = 0$ vs. H_A: $\beta_0 = .2$ was calculated, using the empirical root-mean-squared-error as the standard error of the estimate of the intercept. While in the simulations $\beta_0 = .8$, the power for the test of H_A: $\beta_0 = .8$ is simply too high across all conditions and estimation techniques to be interesting; thus, a lower value, .2 (that of a "low effect size", according to Cohen, 1988) was used. The H_A values used for $\beta_1, \beta_2,$ and β_3 were the true values of the hyperparameters used in the simulations.

Average power values for the test for the intercept are in Tables 7.3. Values of α used are .01 and .05, given the high power of the test. Average power values for the test of the slopes are in Tables 7.4 - 7.6. Values of α used are .05 and .15, given the low power of the tests and the interest researchers might have in marginal statistical significance of predictors' estimates when constructing their models.

Table 7.3

Average Power of H_O: $\beta_0 = 0$ vs. H_A: $\beta_0 = .20$, Across Hyperparameters Known to Meta-Analyst					
				H_A: $\beta_0 = .20$	
				CC	ML
$\alpha = .01$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.281	.548
			Avg. $n_i = 400$	.721	.914
		$k = 100$	Avg. $n_i = 80$	.723	.929
			Avg. $n_i = 400$	.978	.999
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.131	.428
			Avg. $n_i = 400$	.355	.775
		$k = 100$	Avg. $n_i = 80$	.379	.895
			Avg. $n_i = 400$	.837	.996
$\alpha = .05$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.782	.980
			Avg. $n_i = 400$	.992	1.00
		$k = 100$	Avg. $n_i = 80$	.998	1.00
			Avg. $n_i = 400$	1.00	1.00
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.389	.926
			Avg. $n_i = 400$	.842	.994
		$k = 100$	Avg. $n_i = 80$	.903	1.00
			Avg. $n_i = 400$	.999	1.00

Table 7.4

				$H_A: \beta_1 = .222$		$H_A: \beta_1 = .380$	
Average Power of $H_O: \beta_1 = 0$ vs. $\beta_1 = \beta_A$, Across Hyperparameters Known to Meta-Analyst							
				CC	ML	CC	ML
$\alpha = .05$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.057	.070	.095	.159
			Avg. $n_i = 400$	.089	.141	.233	.367
		$k = 100$	Avg. $n_i = 80$	.091	.134	.241	.378
			Avg. $n_i = 400$	.225	.337	.553	.700
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.037	.057	.055	.126
			Avg. $n_i = 400$	.053	.100	.103	.245
		$k = 100$	Avg. $n_i = 80$	.059	.117	.122	.334
			Avg. $n_i = 400$	.118	.272	.322	.618
$\alpha = .15$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.131	.167	.211	.307
			Avg. $n_i = 400$	.200	.278	.394	.540
		$k = 100$	Avg. $n_i = 80$	.204	.271	.407	.553
			Avg. $n_i = 400$	.383	.505	.693	.814
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.104	.143	.139	.260
			Avg. $n_i = 400$	.135	.218	.222	.416
		$k = 100$	Avg. $n_i = 80$	.146	.246	.253	.510
			Avg. $n_i = 400$	.245	.438	.486	.758

Table 7.5

Average Power of H_O: $\beta_2 = 0$ vs. H_A: $\beta_2 = \beta_A$, Across Hyperparameters Known to Meta-Analyst							
				H_A: $\beta_2 = .341$		H_A: $\beta_2 = .518$	
				CC	ML	CC	ML
$\alpha = .05$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.066	.081	.140	.197
			Avg. $n_i = 400$	.142	.181	.366	.474
		$k = 100$	Avg. $n_i = 80$	.143	.178	.377	.481
			Avg. $n_i = 400$	.366	.442	.701	.786
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.044	.058	.072	.119
			Avg. $n_i = 400$	.069	.095	.148	.239
		$k = 100$	Avg. $n_i = 80$	.082	.119	.194	.341
			Avg. $n_i = 400$	.189	.293	.474	.640
$\alpha = .15$	50% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.160	.187	.279	.356
			Avg. $n_i = 400$	.280	.330	.534	.627
		$k = 100$	Avg. $n_i = 80$	.283	.328	.546	.634
			Avg. $n_i = 400$	.303	.598	.813	.872
	75% Incidence of M. Data	$k = 40$	Avg. $n_i = 80$	.118	.146	.172	.249
			Avg. $n_i = 400$	.166	.210	.290	.406
		$k = 100$	Avg. $n_i = 80$	.188	.247	.350	.510
			Avg. $n_i = 400$	.342	.451	.628	.763

Table 7.6

				H_A: β_3 = .619		H_A:β_3 = .656	
Average Power of H_O: β_3 = 0 vs. H_A: β_3 = β_A, Across Hyperparameters Known to Meta-Analyst							
				CC	ML	CC	ML
α = .05	50% Incidence of M. Data	$k = 40$	Avg. n_i = 80	.124	.166	.203	.291
			Avg. n_i = 400	.312	.411	.487	.600
		$k = 100$	Avg. n_i = 80	.326	.401	.512	.615
			Avg. n_i = 400	.650	.727	.814	.880
	75% Incidence of M. Data	$k = 40$	Avg. n_i = 80	.065	.100	.092	.171
			Avg. n_i = 400	.137	.216	.227	.371
		$k = 100$	Avg. n_i = 80	.162	.273	.273	.486
			Avg. n_i = 400	.418	.586	.600	.776
α = .15	50% Incidence of M. Data	$k = 40$	Avg. n_i = 80	.255	.314	.362	.463
			Avg. n_i = 400	.479	.566	.639	.731
		$k = 100$	Avg. n_i = 80	.497	.567	.661	.744
			Avg. n_i = 400	.774	.830	.894	.937
	75% Incidence of M. Data	$k = 40$	Avg. n_i = 80	.158	.219	.206	.323
			Avg. n_i = 400	.274	.374	.389	.544
		$k = 100$	Avg. n_i = 80	.309	.437	.441	.636
			Avg. n_i = 400	.579	.708	.731	.860

A scan of the above Tables allows for some general conclusions. First, except for small meta-analyses with large amounts of missing data and small average study sample sizes, the power of the test of $H_{0:}$ $\beta_0 = 0$ is very, very strong, even when the alternative hypothesis is that β_0 is very small. Second, across all conditions the powers of the slope tests $H_{0:}$ $\beta_j = 0$ ($j = 1$ to 3) are often quite low, though the power is strongly dependent on the size of the slope in the alternative hypothesis. Finally, because of the nature of the normal curve and the nature of power calculations, the superiority of the power of the ML estimates is most obvious when the CC power is neither extremely high or extremely low. If the power for a test of a CC estimator is about .05, it is unlikely that the power for the same test of an ML estimator will be greater than .08. However, in many instances, when the CC power is about .30, the ML power will be .45 or better.

What can a researcher who has a situation with 50% missing data, a large number of studies, and a large average study sample size gather from these tables? First, as long as the data are MCAR in nature, she will get a very precise estimate of β_0 (or, the average effect size, assuming the predictors are grand-mean centered), even if she uses CC estimation. Second, the added power from ML estimation will depend on the size of the effects she is trying to find. If she is trying to find small or moderate effects, EM estimation may give increases in power from 22.5% to 33.7%, or 55.3% to 70.0% (Table 7.3). These figures are for the slope of the completely-observed variable, β_1. For the partially observed variable β_2 the improvements are smaller: from 36.6% to 42.2% and from 70.1% to 78.6% (Table 7.4). The improvement is similar for β_3, except that the average power levels are higher because the larger values of β_3 are examined.

176

The increase in efficiency can thus be reduced to an increase in power. Whether this increase in power is trivial depends on the interpretation, but EM estimation in this instance is a reasonable alternative given the low power of the CC estimates of the slopes. Even for the completely observed variable, the meta-analyst has fairly low power to detect a small slope; however, she is 50% more likely to find one with EM estimation than with CC estimation[5]. She has a moderate chance to detect a medium-sized slope; however, she is 27% more likely to find one with EM estimation than with CC estimation. These small increases in power lead to definite increases in the odds of discovering real relationships between predictors and the outcome. The increases for the partially observed variable β_2, using the power values cited above, are smaller: 15% and 12% increases in the odds of finding significant effects. It would be easier to argue that these improvements do border on trivial.

However, the formula for the calculation of a necessary sample size given a desired α and power level (Mendenhall et al., 1988) leaves us with one last way of measuring "payoff": how many more studies like the ones already gathered would the researcher have to obtain an increase in efficiency comparable to that offered by ML analysis? Using the "worst case scenario" above of 50% missing data, 100 studies, and an average study sample size of 400, the results are as follows:

For the completely observed variable β_1: an increase in power of 22.5% to 33.7% would require 28 more studies, while an increase in power of 55.3% to 70% would require 20 more studies.

[5]The ratio of power values for a small effect is 33.7%/22.5% = 1.50

For the partially observed variable β_2: an increase in power of 36.6% to 42.2% would require 9 more studies, and an increase in power of 70.1% to 78.6% would require 10 more studies.

It is left to the meta-analyst whether the work of locating and coding these additional studies is as costly and time-consuming as learning to apply EM estimation. Suppose, though, that it takes a researcher 200 hours to find, code, and enter 100 studies. (This may be a conservative estimate in some subject areas, a liberal estimate in others). It is fair to assume that most of the findings that the meta-analyst will make depend on the precision of his estimation. Thus, if it takes 200 hours to find and code 100 studies, it is worth 20 hours of work to learn how to conduct a maximum-likelihood estimation of one's data, which would be equivalent to adding 10 studies to estimate those slopes for which there is missing data. Or perhaps it would be worth 40 hours of work to employ ML, as the increase in power to detect an effect for a completely observed variable would be equivalent to that achieved by finding 20 more studies. It should be added that these calculations ignore the large increases in efficiency in the estimation of τ for maximum-likelihood, and the advantages of ML estimation in handling different missing-data mechanisms. It also might be argued that these calculations ignore the many hours spent on primary studies, hours that are to some degree wasted when the most efficient estimation techniques are not employed.

4. Future Research

As noted above, further simulation work (though perhaps not as intensive as what
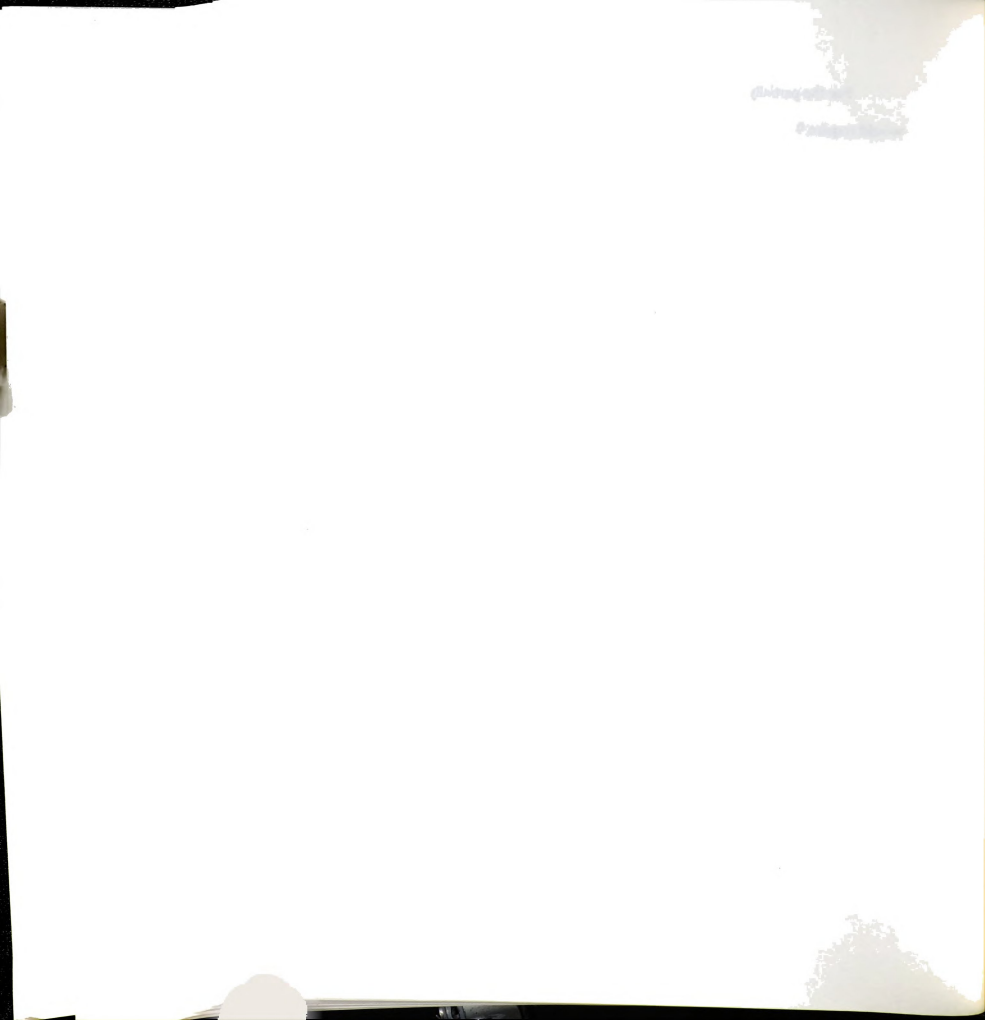
was conducted here) is needed to confirm the superiority of ML estimation when there are many predictors. More important is an investigation into whether the current estimation procedure can handle dichotomous predictors that have missing data. While the brief investigation in Chapter V, Section 7 suggests that the ML estimation derived here is robust to the assumption of normality of predictors, further investigation is in order. If ML estimation is not as robust as expected for dichotomous predictors, a ML method for meta-analytic data that considers a multinomial distribution for categorical predictors should be derived. Within this investigation, the accuracy of bootstrapped standard errors for data with dichotomous predictors should be considered. Similarly, it may be worthwhile to investigate the accuracy of bootstrapped standard errors for different missing-data mechanisms, though the work of Su (1988) suggests that such standard errors should still be accurate. An investigation into how this ML estimation procedure might handle measurement error in the predictors would also be worthwhile.

Finally, it would be beneficial to derive a maximum-likelihood method that could handle dependent effects (such as the analysis of Moritz et al., 2000). This step adds a layer of complexity to an already complex model; however, many meta-analysts must find some way to deal with dependent effects and it would be ideal for a maximum-likelihood estimation program to have the ability to handle them.

5. Conclusion

The purpose of this research was to investigate alternatives to complete-case estimation of meta-analytic parameters, with a focus on maximum-likelihood estimation

179

but with consideration of available-case estimation as a simpler and perhaps substantively equivalent alternative. A maximum-likelihood method designed to handle the intricacies of meta-analytic data was derived and tested through simulation in Chapter V and applied to a real dataset in Chapter VI. Analyses in Chapters V and VII suggest that ML methods lead to important gains in efficiency that are generally superior to those found through AC methods. However, the gains in efficiency are not always great, and depend on many hyperparameters, some of them easily discernable to the meta-analyst, such as number of studies, amount of missing data, and average study sample size. Relative efficiencies were lowest for large numbers of studies with large sample sizes and little missing data. Even in this case, however, the "payoff" of ML estimation may be acceptable relative to the time spent with the many other tasks associated with conducting a meta-analysis. When other factors are considered, such as the clearly superior efficiency of ML estimation of τ and the ability of ML estimation to handle different types of missing data, the payoff is even greater.

The "payoff" is not acceptable, of course, if there is no accessible software for the meta-analyst to run. Creation of a program that conducts maximum-likelihood estimation takes a considerable amount of time. However, I am making the program used to analyze the Lipsey dataset in Chapter VI available[6]. The program runs in SAS/IML, accepts ASCII data, and can be learned in under an hour if one is even slightly familiar with SAS. In addition, little "customization" is needed to make the program able to handle any given

---

[6]   The author may be contacted at fahrbach@msu.edu or through Dr. Betsy Becker, Erickson Hall, Michigan State University, East Lansing, MI. (517) 355-9567.

dataset. With this program available, the main hurdle to maximum-likelihood analysis of meta-analytic data is overcome. This research moves maximum-likelihood estimation from merely "something to be considered" to a method that, across a wide range of realistic conditions, must be considered strongly advisable.
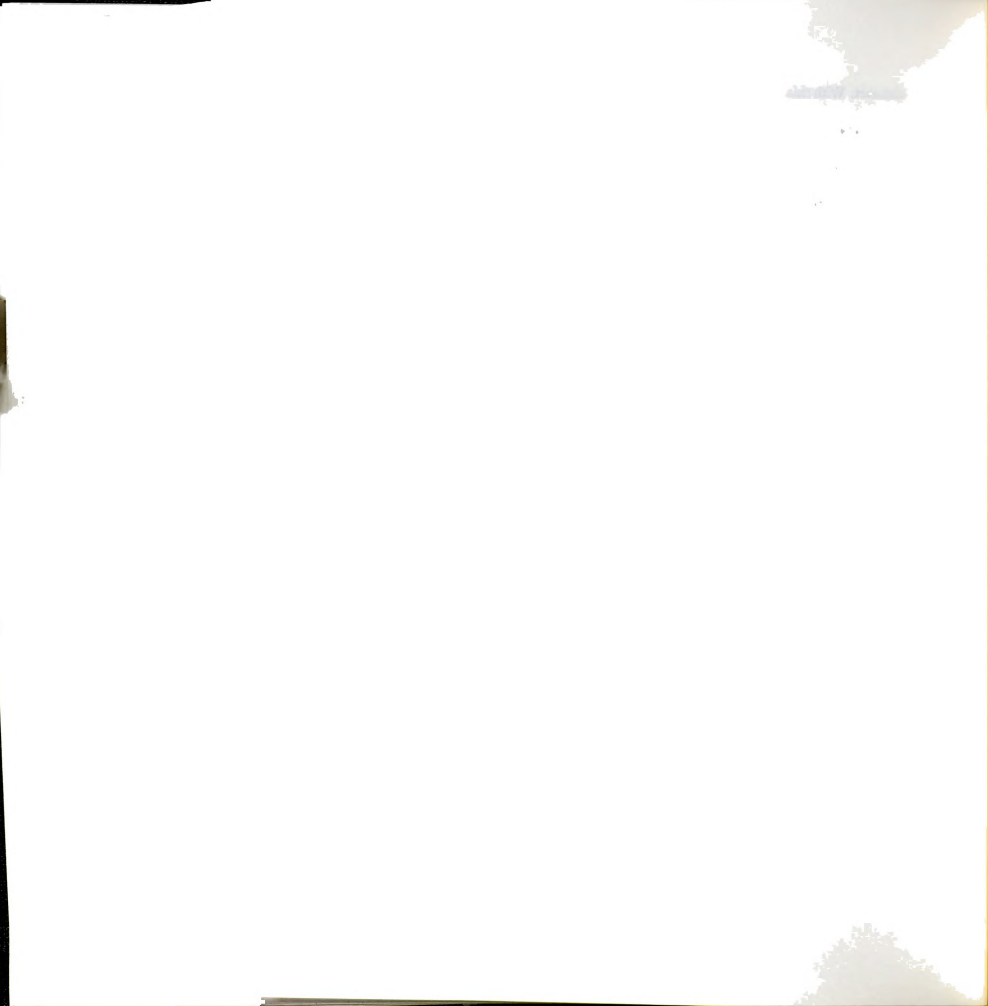
APPENDIX

Study Sample Sizes and Missing Data Patterns

There were eight combinations of k, average study sample size, and incidence of missing data. The tables below detail the distribution of missing data on the 2nd and 3rd predictors for these eight combinations of hyperparameters. Each cell represents one study. The first number in a cell is a study sample size, while the second refers to a pattern of missing data. The label "C" means that data were complete: the study had observations on all three predictors. The label "M2" means that only the 2nd predictor was missing, "M3" means that only the 3rd predictor was missing, and "M23" means that both the 2nd and 3rd predictors were missing.

Thus, within each combination of hyperparameters, the patterns of missingness, the average study sample size, and the mapping of the missing data patterns to particular groups of study sample sizes are all held constant. This was done in order to limit the amount of variance in the findings to as few sources as possible (i.e., the seven hyperparameters). Future simulation work might allow more flexibility in how study sample sizes are assigned and allow for missing data patterns that vary between simulated meta-analyses.

Table A.1

$k = 40$, Average $n_i = 80$, 50% Incidence of Missing Data			
40,M23	160,M2	40,C	40,C
40,M23	320,M2	40,C	40,C
40,M23	40,M3	40,C	40,C
160,M23	40,M3	40,C	40,C
40,M2	40,M3	40,C	40,C
40,M2	40,M3	40,C	40,C
40,M2	40,M3	40,C	40,C
40,M2	40,M3	40,C	40,C
40,M2	160,M3	160,C	160,C
40,M2	320,M3	320,C	320,C

Table A.2

$k = 40$, Average $n_i = 80$, 75% Incidence of Missing Data			
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
40,M23	40,M2	40,M3	40,C
160,M23	160,M2	160,M3	160,C
320,M23	320,M2	320,M3	320,C

Table A.3

$k = 40$, Average $n_i = 400$, 50% Incidence of Missing Data			
80,M23	800,M2	80,C	80,C
80,M23	1600,M2	80,C	80,C
320,M23	80,M3	80,C	80,C
800,M23	80,M3	80,C	80,C
80,M2	80,M3	320,C	320,C
80,M2	320,M3	320,C	320,C
80,M2	320,M3	320,C	320,C
320,M2	320,M3	320,C	320,C
320,M2	320,M3	800,C	800,C
320,M2	1600,M3	1600,C	1600,C

Table A.4

$k = 40$, Average $n_i = 400$, 75% Incidence of Missing Data			
80,M23	80,M2	80,M3	80,C
80,M23	80,M2	80,M3	80,C
80,M23	80,M2	80,M3	80,C
80,M23	80,M2	80,M3	80,C
320,M23	320,M2	320,M3	320,C
320,M23	320,M2	320,M3	320,C
320,M23	320,M2	320,M3	320,C
320,M23	320,M2	320,M3	320,C
800,M23	800,M2	800,M3	800,C
1600,M23	1600,M2	1600,M3	1600,C

Table A.5

k = 100, Average n_i = 80, 50% Incidence of Missing Data				
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
40,M23	40,M2	40,M3	40,C	40,C
160,M23	160,M2	160,M3	160,C	160,C
320,M23	320,M2	320,M3	320,C	320,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
40,M2	40,M3	40,C	40,C	40,C
160,M2	160,M3	160,C	160,C	160,C
320,M2	320,M3	320,C	320,C	320,C

185

Table A.6

$k = 100$, Average $n_i = 80$, 75% Incidence of Missing Data
40,M23	160,M23	40,M2	40,M3	40,C
40,M23	320,M23	40,M2	40,M3	40,C
40,M23	160,M23	40,M2	40,M3	40,C
40,M23	320,M23	40,M2	40,M3	40,C
40,M23	160,M23	40,M2	40,M3	40,C
40,M23	40,M2	320,M2	40,M3	40,C
40,M23	40,M2	160,M2	40,M3	40,C
40,M23	40,M2	320,M2	40,M3	40,C
40,M23	40,M2	160,M2	40,M3	40,C
40,M23	40,M2	320,M2	40,M3	40,C
40,M23	40,M2	40,M3	320,M3	40,C
40,M23	40,M2	40,M3	160,M3	40,C
40,M23	40,M2	40,M3	320,M3	40,C
40,M23	40,M2	40,M3	160,M3	40,C
40,M23	40,M2	40,M3	320,M3	40,C
40,M23	40,M2	40,M3	40,C	160,C
40,M23	40,M2	40,M3	40,C	320,C
40,M23	40,M2	40,M3	40,C	160,C
40,M23	40,M2	40,M3	40,C	320,C
40,M23	40,M2	40,M3	40,C	160,C

Table A.7

k = 100, Average n_i = 400, 50% Incidence of Missing Data				
80,M23	80,M2	80,M3	80,C	80,C
80,M23	80,M2	80,M3	80,C	80,C
80,M23	80,M2	80,M3	80,C	80,C
80,M23	80,M2	80,M3	80,C	80,C
320,M23	320,M2	320,M3	320,C	320,C
320,M23	320,M2	320,M3	320,C	320,C
320,M23	320,M2	320,M3	320,C	320,C
320,M23	320,M2	320,M3	320,C	320,C
800,M23	800,M2	800,M3	800,C	800,C
1600,M23	1600,M2	1600,M3	1600,C	1600,C
80,M2	80,M3	80,C	80,C	80,C
80,M2	80,M3	80,C	80,C	80,C
80,M2	80,M3	80,C	80,C	80,C
80,M2	80,M3	80,C	80,C	80,C
320,M2	320,M3	320,C	320,C	320,C
320,M2	320,M3	320,C	320,C	320,C
320,M2	320,M3	320,C	320,C	320,C
320,M2	320,M3	320,C	320,C	320,C
800,M2	800,M3	800,C	800,C	800,C
1600,M2	1600,M3	1600,C	1600,C	1600,C

Table A.8

k = 100, Average n_i = 400, 75% Incidence of Missing Data				
80,M23	800,M23	320,M2	320,M3	80,C
80,M23	1600,M23	320,M2	320,M3	80,C
80,M23	800,M23	320,M2	320,M3	80,C
80,M23	1600,M23	320,M2	320,M3	80,C
80,M23	800,M23	320,M2	320,M3	80,C
80,M23	80,M2	1600,M2	320,M3	320,C
80,M23	80,M2	800,M2	320,M3	320,C
80,M23	80,M2	1600,M2	320,M3	320,C
80,M23	80,M2	800,M2	320,M3	320,C
80,M23	80,M2	1600,M2	320,M3	320,C
320,M23	80,M2	80,M3	1600,M3	320,C
320,M23	80,M2	80,M3	800,M3	320,C
320,M23	80,M2	80,M3	1600,M3	320,C
320,M23	80,M2	80,M3	800,M3	320,C
320,M23	80,M2	80,M3	1600,M3	320,C
320,M23	320,M2	80,M3	80,C	800,C
320,M23	320,M2	80,M3	80,C	1600,C
320,M23	320,M2	80,M3	80,C	800,C
320,M23	320,M2	80,M3	80,C	1600,C
320,M23	320,M2	80,M3	80,C	800,C

REFERENCES

Anderson, A., Basilevsky, A. & Hum, D. (1983). Missing data: A review of the literature. In J.D. Wright, P.H. Rossi, & A.B. Anderson (Ed.), *Handbook of survey research*. New York: Academic Press.

Azen, S. & Van Guilder, M. (1981). Conclusions regarding algorithms for handling incomplete data. *Proceedings of the Statistical Computing Section, American Statistical Association 1981*, 53-56.

Beale, E. & Little, R. (1975). Missing values in multivariate analysis. *Journal of the Royal Statistical Society, B37*, 129-145.

Beaton, A. E. (1964). The use of special matrix operations in statistical calculus. Educational Testing Service Research Bulletin, RB-64-51.

Becker, B. J. (1985). *Tests of combined significance: Hypotheses and power considerations*. Unpublished doctoral dissertation, University of Chicago.

Begg, C. B. (1994). Publication bias. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis*. New York: Russell Sage Foundation.

Bryk, A. S. & Raudenbush, S. W. (1992). Hierarchical linear models: applications and data analysis methods. Newbury Park, CA: Sage Publications, Inc.

Buck, S. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society, B22*, 302-306.
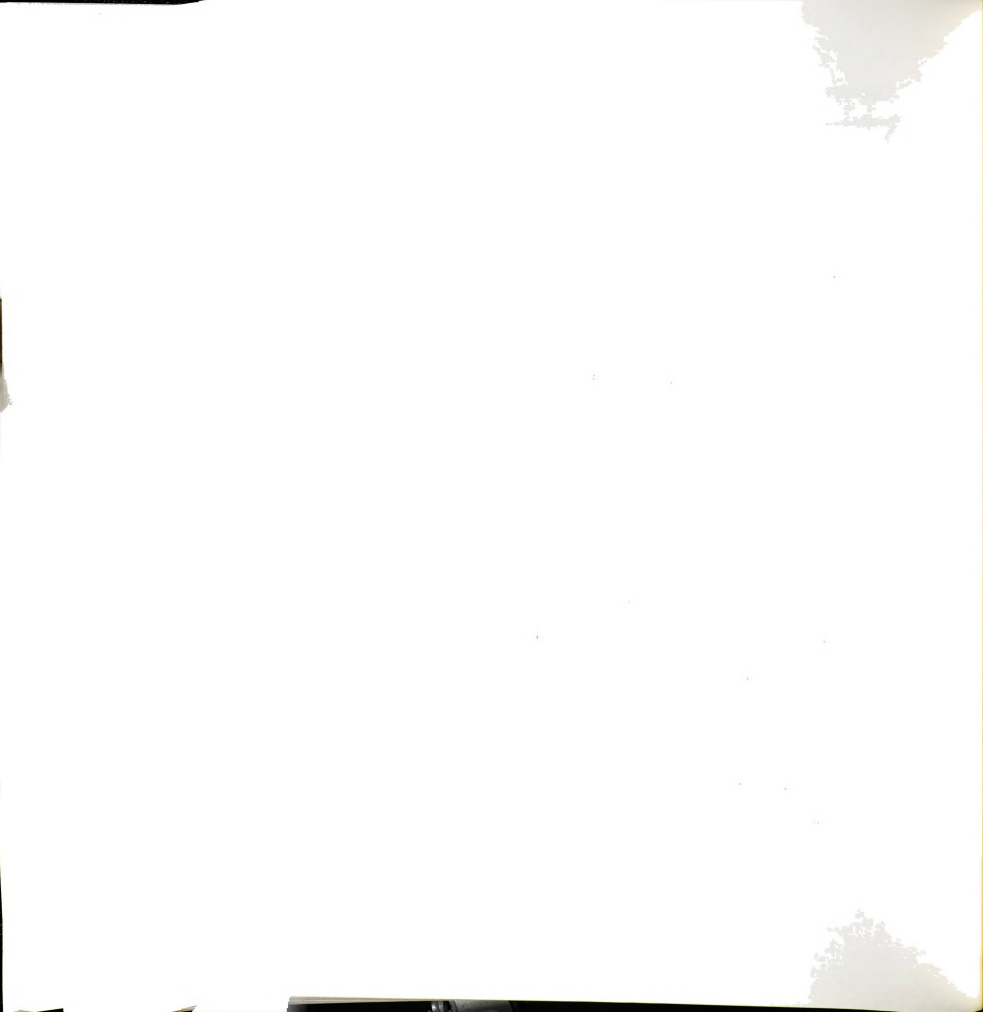
Chang, L. (1992). *A power analysis of the test of homogeneity in effect-size meta-analysis*. Unpublished doctoral dissertation, Michigan State University.

Cohen, J. (1988). *Statistical power analyses for the behavioral sciences* (2nd ed.) Hillsdale, NJ: Erlbaum.

Cook, R., Tsai, C., & Wei, B. (1986). Bias in nonlinear regression. *Biometrika, 73*(3), 615-623.

Cordeiro, G. & McCullagh, P. (1991). Bias correction in generalized linear models. *Journal of the Royal Statistical Society B, 3*, 629-643.

Dempster, A.P, Laird, N.M., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society (with discussion), B39*, 1-38.

Dixon, W. J. (1992). *BMDP Statistical Software Manual (Vol. 2)*. Berkeley, CA: University of California Press.

Efron, B. & Tibshirani, R. (1991). Statistical data analysis in the computer age. *Science, 253*, 390-395.

Fahrbach, K. (1995). *A Monte-Carlo investigation of univariate and multivariate meta-analysis*. Unpublished apprenticeship paper, Michigan State University.

Gay, L. R. (1992). *Educational research: competencies for analysis and application*. New York: Merrill.

Glynn, R., Laird, N., & Rubin, D. B. (1986). Selection modeling versus mixture modeling with nonignorable nonresponse. In H. Wainer (Ed.), *Drawing inferences from self-selected samples*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
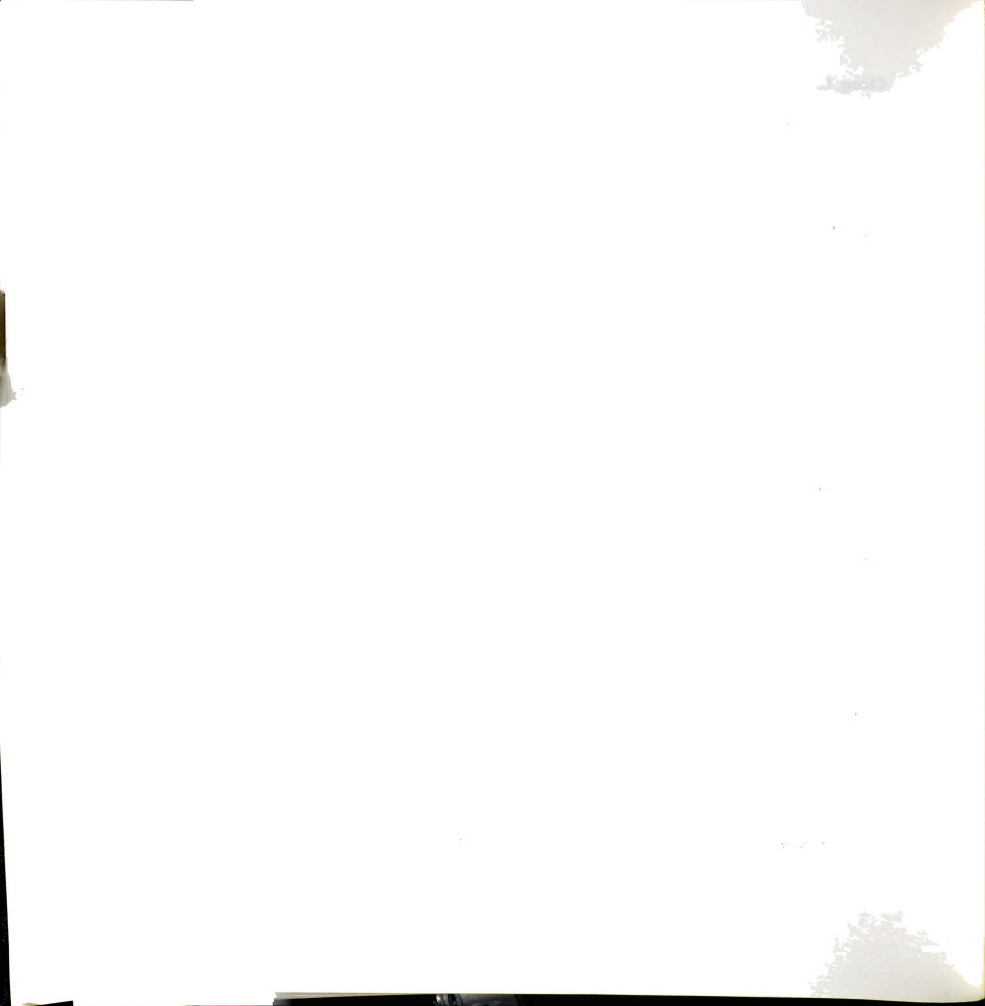
Green, B. & Hall, J. (1984). Quantitative methods for literature review. *Annual Review of Psychology, 35*, 37-53.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society ,B30*, 67-82.

Harris, M. & Rosenthal, R. (1985). Mediation of interpersonal expectancy effects: 31 meta-analyses. *Psychological Bulletin, 97(3)*, 363-386.

Hunter, J. E. & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Beverly Hills, CA: Sage.

Kim, J. & Curry, J. (1977). The treatment of missing data in multivariate analysis. *Sociological Methods and Research, 6*, 215-240.

Lent, R. H., Auerbach, H. A., & Levin, L. S. (1971). Research design and validity assessment. *Personnel Psychology, 24*, 247-274.

Lipsey, M. W. (1999a). Can intervention rehabilitate serious delinquents? *American Academy of Political and Social Science, 564*, 142-166.

Lipsey, M. W. (1999b). Can rehabilitative programs reduce the recidivism of juvenile offenders? *Virginia Journal of Social Policy and the Law, 6*(3), 611-641.

Little, R. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistician, 37*, Vol.1, 23-28.

Little, R. (1992). Regression with missing X's: A review. *Journal of the American Statistical Association, 87*, 1227-1237.

Little, R. (2000). *Personal communication*, August, 2000.

Little, R.J.A. and Raghunathan, T. E. (1999). On summary-measures analysis
of the linear mixed-effects model for repeated measures when data are not
missing completely at random. *Statistics in Medicine, 18*, 2465-2478.

Little, R. & Rubin, D. (1987). *Statistical analysis with missing data.* New York: John
Wiley.

Matthai, A. (1951). Estimation of parameters from incomplete data with application to
design of sample surveys. *Sankhya, 2*, 145-152.

Mendenhall, W., Scheaffer, R., & Wackerly, D. (1986). *Mathematical Statistics with
Applications.* Duxbury Press: Boston.

Meng, X. & Rubin, D. (1991). Using EM to obtain asymptotic variance-covariance
matrices: The SEM algorithm. *Journal of the American Statistical Association,
86*, 899-909.

Morrison, D. F. (1967). *Multivariate statistical methods.* New York: McGraw-Hill.

Moritz, S., Feltz, D., Fahrbach, K., & Mack, D. (2000). The relation of self-efficacy
measures to sport performance: A meta-analytic review. *Research Quarterly for
Exercise and Sport, 71*(3), 280-291.

Pigott, T. (1992). *The application of normal maximum-likelihood methods to missing
data in meta-analysis.* Unpublished doctoral dissertation, University of Chicago,
Chicago IL.

Pigott, T.(1994). Methods for handling missing data in research synthesis. In H. Cooper &
L. V. Hedges (Eds.), *Handbook of research synthesis.* New York: Russell Sage
Foundation.

Raudenbush, S. W. (1994). Random effects models. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis*. New York: Russell Sage Foundation.

Schafer, J. (1997a). *Analysis of incomplete multivariate data*. New York: Chapman & Hall.

Schafer, J. (1997b). Imputation of missing covariates under a multivariate linear mixed model. Technical Report, Department of Statistics, Pennsylvania State University

Schafer, J. (1998). Some improved procedures for fixed linear models. Technical Report, Department of Statistics, Pennsylvania State University.

Searle, S.R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: John Wiley & Sons.

Shadish, W. R. & Haddock, C. K. (1994). Combining estimates of effect size. In H. Cooper & L. V. Hedges (Eds.), *Handbook of research synthesis*. New York: Russell Sage Foundation.

Slavin, R. E. (1984). Meta-analysis in education: How has it been used? *Educational Researcher, 13*(8), 6-15.

Su, H. (1988). *Estimation of standard errors in some multivariate models with missing data*. Unpublished doctoral dissertation, University of Michigan.

Tatsuoka, M. (1988). *Multivariate analysis.* .New York: Macmillan.

Van Praag, B.M.S, Dijkstra T.K., & Van Veltzen, J. (1985). Least squares theory based on distributional assumptions with an application to the incomplete observations problem. *Psychometrika, 50*, 25-36.