


SEVEN METHODS OF HANDLING MISSING DATA
USING SAMPLES FROM A NATIONAL DATA BASE

by


Eleanor Lea Witta

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY
in
Educational Research and Evaluation

APPROVED:




J. Kaiser, Chairman



J. Fortune



K. Hereford



T. Z. Keith



D. Strickland

c.2

LD
5655
Y856
1992
W588
c.2

SEVEN METHODS OF HANDLING MISSING DATA
USING SAMPLES FROM A NATIONAL DATA BASE

by

Eleanor Lea Witta

Committee Chairman: Javaid Kaiser
Educational Research and Evaluation

(ABSTRACT)

The effectiveness of seven methods of handling missing data was investigated in a factorial design using random samples selected from the National Education Longitudinal Study of 1988 (NELS-88). Methods evaluated were listwise deletion, pairwise deletion, mean substitution, Buck's procedure, mean regression, one iteration regression, and iterative regression. Factors controlled were number of variables (4 and 8), average intercorrelation (0.2 and 0.4), sample size (200 and 2000), and proportion of incomplete cases (10%, 20%, and 40%). The pattern of missing values was determined by the pattern existing in the variables selected from NELS-88 data base.

Covariance matrices resulting from the use of each missing data method were compared to the 'true' covariance matrix using multi-sample analysis in LISREL 7. Variable means were compared to the 'true' means using the MANOVA procedure in SPSS/PC+. Statistically significant differences ($p < .05$) were detected in both comparisons.

The most surprising result of this study was the effectiveness ($p > .05$) of pairwise deletion whenever the sample size was large thus supporting the contention that the error term disappears as sample size approaches infinity (Glasser, 1964). Listwise deletion was also effective ($p > .05$) whenever there were four variables or the sample size was small. Almost as surprising was the relative ineffectiveness ($p < .05$) of the regression methods. This is explained by the difference in proportion of incomplete cases versus the proportion of missing values, and by the distribution of the missing values within the incomplete cases.

Acknowledgement

First I would like to recognize the support provided by my family. My children, Lori, Christy, and Benjamin, were encouraging when things were tough and were accepting when I had to be absent. Only those who have been in a similar position can comprehend the importance of their support. But most of all, my husband George, has provided support and confidence when I felt this task was impossible. It was he who set the example of acceptance for our children.

I would also like to acknowledge the contribution of my committee and the staff of the department of educational research. Although it is within their job description to assist in preparation of a student's dissertation, some have done far more than the job requires. Jimmie Fortune has been supportive and provided practical advice from the day I entered doctoral candidacy. Tim Keith has proved to be truly a "mentor extraordinaire". Deborah Strickland first introduced me to secondary data bases and the use of mainframe facilities. Lawrence Cross reviewed and corrected many revisions of the prospectus for this document. Carmen Wisdom kept me current with time schedules, and, when I could not commute to school to complete required forms, she saw that they were completed. The special help these individuals have provided is not within their job

descriptions. It simply reflects their concern for a student.

Most of all, I am grateful that God has permitted me to live in a time when females may participate in this program and has provided me with the support of individuals who were essential to its completion. Some of these individuals I may not see again, so I shall simply thank God for the privilege of meeting them at all.

Table of Contents

| Chapter | Title | Page |
|-------------|--|------|
| I | Introduction | 1 |
| | Statement of the Problem | 3 |
| | Significance of the Study | 7 |
| | Statement of Hypothesis | 8 |
| II | Review of Literature | 9 |
| III | Method. | 26 |
| IV | Results | 40 |
| V | Discussion and Conclusions | |
| | Discussion | 52 |
| | Conclusions | 64 |
| | References | 66 |
| Appendix A: | | |
| | Missing Data Characteristics | 73 |

List of Tables

| Table | Title | Page |
|------------|---|------|
| 1 | Correlation Matrices and Descriptive Statistics of Four Variables | 30 |
| 2 | Correlation Matrices and Descriptive Statistics of Eight Variables | 33 |
| 3 | Comparison of Proportion of Incomplete Cases to Proportion of Missing Values | 36 |
| 4 | Results of Retaining Four-Variable Covariance Matrix | 41 |
| 5 | Results of Retaining Eight-Variable Covariance Matrix | 45 |
| 6 | Results of Four-Variable Means Comparison | 47 |
| 7 | Results of Eight-Variable Means Comparison | 48 |
| 8 | Summary Table Identifying Methods That Successfully Reproduced the Target Sample Covariance Matrix in the Experimental Conditions | 50 |
| 9 | Summary Table Identifying Methods That Successfully Reproduced the Target Sample Means in the Experimental Conditions | 51 |
| Appendix A | | |
| 1 | Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Four Variables | 74 |
| 2 | Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables | 75 |
| 3 | Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Four Variables | 79 |
| 4 | Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Eight Variables | 80 |

CHAPTER I

Introduction

When data is analyzed in survey research, often there are missing values. If the mechanism causing the missing values is known, the solution to this problem may be incorporated in the study. Many times, however, the mechanism causing the missing values is not known. Ignoring this problem may lead to analysis of data that is of dubious value. Publication of the results of this analysis without correctly handling the missing values may "jeopardize the credibility of the organization conducting the survey and preparing the analysis and report:..." (Little & Smith, 1983, p. 518). Unfortunately, there is no established correct method for handling missing values when the mechanism causing them is unknown.

In addition, different methods of handling missing values may produce different results. When Jackson (1968) entered data on all the available variables in a discriminant analysis, the significance of the regression coefficients of individual variables, as well as the interpretation of the importance of these variables, changed with the missing value method used. Witta and Kaiser (1991) also reported that the regression coefficients and total variance accounted for by the variables changed depending on the method used to handle missing values. After reanalyzing

three studies of private/public school achievement, Ward and Clark III (1991) concluded that the method used to handle missing data influenced the outcome of these studies.

The effects of different methods of handling missing values are studied by data simulation and the use of data from studies that contain missing values. Data simulation uses a correlation matrix either from a real study (Timm, 1970) or designed by the researcher to simulate planned characteristics (Gleason & Staelin, 1975). The researcher designed simulations are chosen to include a specific number of variables and average intercorrelation. Samples of a predetermined size are generated from this correlation matrix. Missing values are then created in the simulated data set either by random deletion of values (Chan & Dunn, 1972; Gleason & Staelin, 1975; Kaiser & Tracy, 1988; Timm, 1970) or by systematic deletion (Haitovsky, 1968; Kaiser, 1990). Unfortunately, values in the behavioral sciences are seldom randomly missing (Cohen & Cohen, 1983) and the studies using systematic deletion have provided no rationale for the systematic method used.

Since the true estimates for the values that are missing in real data sets are not known, an attempt is made to predict a known variable after treating data by a missing data handling method. Prediction has been accomplished by discriminant analysis (Jackson, 1968) or regression (Witta &

Kaiser, 1991). Since the true values of the missing data are not known in these studies, it is difficult to determine which missing data method was more effective. In addition, studies conducted on real data usually report neither the correlation matrix nor the pattern of missing values, thus hampering attempts to synthesize their findings.

Consequently, studies that use simulated data are perceived as more generalizable to real data situations than studies conducted on real data.

Both of these approaches offer advantages. The study of real data with naturally occurring missing values offers direct applicability. The use of data simulation provides the opportunity to manipulate conditions under which the missing values are studied. In the past, methods of handling missing data have been studied by data simulation under conditions determined by average intercorrelation, number of variables, sample size, proportion of missing values, and the pattern of missing data (Gleason & Staelin, 1975; Kaiser & Tracy, 1988). This study was designed to incorporate the perceived strengths of each of these approaches.

Statement of the Problem

The purpose of this study was to investigate the effectiveness of seven methods of handling missing data.

Effectiveness was defined as the probability of accurately reproducing the true covariance matrix or variable means.

The methods studied were listwise deletion, pairwise deletion, mean substitution, Buck's procedure, mean regression, 1-iteration regression, and iterative regression. Listwise deletion removes any case with one or more missing values from analysis. Pairwise deletion computes covariances using pairs that have both observations. Mean substitution imputes the variable mean to replace any missing value. The initial correlation matrix for Buck's procedure is produced by using the listwise deletion method. The resulting matrix is used to develop regression equations and to produce estimates of the missing values. The variable mean is substituted for missing values to produce the initial correlation matrix for the mean regression method. Then each variable with a missing value is regressed on the remaining variables. The new estimate is obtained and it replaces the variable mean initially substituted for the missing value. 1-iteration regression initiates itself from the results of the mean regression method and terminates with another round of regression. Each variable with a missing value (previously replaced by an estimate produced by the mean regression method) is regressed on all remaining variables to produce new estimates. These estimates were obtained and replaced

the prior ones. Iterative regression is the continuation of the one-iteration method. The development of new regression equations and estimates is continued until subsequent iterations do not produce a change in the variable means or standard deviations or when a pre-specified number of iterations is completed.

A factorial design using two levels each of average intercorrelation, sample size, and number of variables and three levels of the proportion of incomplete cases was used to test these methods. The number of variables (4 and 8) were chosen arbitrarily as multiple indicators. Four variables were also used by Buck (1960), Beale and Little (1975), Chan and Dunn (1972), Chan, Gilman and Dunn (1976), and Haitovsky (1968). Eight variables have previously been used by Chan and Dunn (1972), Chan et al. (1976), and Timm (1970). Use of two different levels of variables provides the opportunity to evaluate the effectiveness of each missing data treatment method at each level.

The average intercorrelation was determined by Kaiser's Gamma (1962) and is computed by the expression:

$$\gamma = (\underline{\lambda} - 1) / (\underline{p} - 1)$$

where $\underline{\lambda}$ is the largest eigenvalue and \underline{p} is the number of variables. The average intercorrelation produced by Kaiser's Gamma is approximately equal to the mean of the absolute values of zero-order correlations among the

variables. The two levels of average intercorrelation selected for this study were 0.2 and 0.4. Prior studies that have included these levels of average intercorrelation include Timm (1970), Gleason and Staelin (1975), Chan and Dunn (1972), and Chan et al. (1976). Of these, only Timm's (1970) study used correlation matrices from previous research. The remainder generated correlation matrices according to the specifications of their study. Other studies that have used this factor to determine the data matrix generated include Kaiser (1983, 1988).

Sample sizes of 200 and 2000 were selected for this study. A sample size of 200 has previously been used by Beale and Little (1975), Gleason and Staelin (1975), and Timm (1970). No other study was found that used a sample size of 2000. However, Haitovsky (1968) and Kim and Curry (1977) used 1000; Greenlees, Reece, and Ziechang (1982) used 5,364; Jackson (1968) used 14,693; and Little and Su (1987) used 4,764.

The proportion of incomplete cases was studied using 10%, 20%, and 40% incomplete cases. An incomplete case was defined as a case having one or more missing values on the selected variables. Most of the data simulation studies reviewed have used proportion of missing values rather than proportion of incomplete cases. The proportion of missing values was defined as the total number of missing values on

all variables divided by the total number of data points in the sample. To provide for a comparison, the proportion of missing values in this study was also calculated.

Significance of the Study

The variables used in this study have been selected from the parent and student supplements of the National Education Longitudinal Study of 1988 (NELS-88). Since this is a relatively new secondary data base, the results from this study about handling missing values will be of use to future studies conducted using this data base. The variables used in this study were also chosen from those currently being used in a continuing study of parental involvement at Virginia Polytechnic Institute and State University (Keith, Bickley, Keith, Trivett, Singh, & Troutman, 1992). Results from this study are directly applicable to the parental involvement study.

In addition, in the area of educational research, listwise and pairwise deletion are used frequently to handle missing values (Keith et al. 1992; Page & Keith, 1981; Wolfle, 1985). Yet Little and Rubin (1987) found it difficult to recommend either of these methods because their performance is unreliable, they may need ad hoc adjustments, and it is difficult to predict when the methods would fail. If these methods are not effective when used with the NELS-88 data base, researchers using this data base need to know.

Although others have studied some of these methods under similar conditions (Chan & Dunn, 1972; Chan et al., 1976; Gleason & Staelin, 1975; Kaiser, 1990; Timm, 1970) none of them have used real data. Nor was any study found that used all of the methods included in this study.

Statement of Hypothesis

The effectiveness of missing data methods will be explored under various experimental conditions determined by the sample size, number of variables, average intercorrelation among variables, and the proportion of incomplete cases.

Limitations

This study is limited in the number of methods compared, the variables selected, and the number of levels of the determining factors. All variables were assumed to have a linear relationship. Since the study was conducted using a single sample for each experimental condition, the standard error is not known. Therefore, the results may not stay the same in future replications.

CHAPTER II

Review of Literature

Gleason and Staelin (1975) divided alternatives for estimating missing data into two categories: the statistical procedure and the pragmatic approach. The statistical procedure assumes the observed data is a sample drawn from a multivariate distribution of known form but unknown parameter. This provides a model of the estimation process for analytic evaluation. To escape these distributional assumptions, the pragmatic approach uses information from the variable without missing values to construct estimation for missing entries. Heuristics guide this process. Evaluation measures how well different methods can reconstruct unknown values from available data using real examples rather than studying the properties of the model. Only methods using the pragmatic approach are considered in this study.

Various methods known to handle missing data and the factors that influence them are discussed in the following section.

MethodsListwise Deletion

Listwise deletion is probably the most frequently used method of handling missing data and is available as a default option in several statistical software including

LISREL, SPSS, NCSS. This method discards cases with a missing value on any variable and thus is very wasteful of data. If the data are assumed to be missing completely at random, however, this method is unbiased and the covariance matrix will not differ from the one formed if none of the values had been missing (Anderson, Basilevsky, & Hum, 1983). Nevertheless, the loss of cases results in a loss of error degrees of freedom yielding a loss of statistical power and a larger standard error (Cohen & Cohen, 1983).

The problem is "... the number of deleted cases increases as the pattern of missing information becomes more random." (Kim & Curry, 1977, p. 216). As a higher number of cases is deleted, the assumption that the remaining data constitutes a random sample becomes less tenable (Hertel, 1976). If the values cannot be assumed to be randomly missing, this method may eliminate entire subgroups who have refused to respond on a single variable. If Jackson (1968) had used this method, she would have eliminated more than 7,000 of the 14,000 cases in her study. The dropped cases would have differed in important characteristics from those retained.

In addition, if the proportion of missing values is higher and the number of cases is not large enough, randomness of the missing values is questionable (Cohen & Cohen, 1983). If the missing data is not random or non-

respondents differ from the respondents on the variable of interest, this method is unsatisfactory (Greenlees et al., 1982).

In a Monte Carlo study, Haitovsky (1968) found listwise deletion superior to pairwise deletion in terms of efficiency and bias of the partial regression coefficients. This study included both random and systematic deletion of values, two to five variables, and large sample sizes (400-1000). Pairwise deletion was superior to listwise in only one instance - when the deletion pattern left only 9-10 percent of the observations available for use in listwise.

In a computer simulation comparison Chan et al. (1976) found listwise deletion better than pairwise deletion, mean substitution, and Buck's regression procedure in correctly classifying cases by discriminant function when there were: (1) two predictors and the determinant function of the correlation matrix was of small or medium size, (2) four predictors and the determinant function was small, and (3) all predictors when the correlation matrix was near singular. For other conditions, listwise deletion was worse than any method except pairwise deletion. This study randomly deleted data for missing values and used only small sample sizes (15-35). In another computer simulation, Timm (1970) found listwise deletion superior to mean substitution

and Buck's regression procedure but only when there were two variables and the average intercorrelation was low.

Listwise deletion performs better when the average intercorrelation is small, the number of independent variables is less than four, and the proportion of missing values is small (Chan et al, 1976; Haitovsky, 1968; Timm, 1970). The assumption of missing completely at random is crucial to the use of this method. If this assumption is satisfied, listwise deletion is unbiased. It is, however, common to find the complete sample differ in important ways from the incomplete sample (Little & Rubin, 1987).

Pairwise Deletion

When using pairwise deletion, covariances are computed between all pairs of variables having both observations, eliminating those that have a missing value for one of the two variables (Glasser, 1964). Means and variances are computed on all available observations. The assumption made is that the use of the maximum number of pairs and all the individual observations yield more valid estimates of the relationship between the variables. The estimates of means and variances produced are more satisfactory than if any available data were excluded (Anderson et al., 1983). It is assumed that when two variables are correlated, information on one improves the estimates of the other variable. It is also assumed that the pairs are a random subset of the

sample pairs. If these assumptions are true, pairwise deletion produces unbiased estimates of the variable means and variances (Hertel, 1976).

When missing data are not missing completely at random, the correlation matrix produced by pairwise deletion may not be Gramian (Norusis, 1988b). A non-Gramian matrix is not symmetric and contains at least one negative eigenvalue. Since total variance is equal to the number of variables, when one eigenvalue is negative, the other eigenvalues are inflated to compensate. This may cause artificially larger factor loadings in factor analysis thus biasing results (Rummel, 1970). Because a correlation matrix developed using this method may not be positive definite, it cannot be inverted (Cohen & Cohen, 1983; Kim & Curry, 1977).

In a study using real data, Buck (1960) found pairwise deletion better than listwise deletion in estimating the variable means and standard deviations using four variables with low average intercorrelation, 12% missing values, and a sample size of 72. Kim and Curry (1977) also found pairwise deletion better than listwise deletion in computing the regression coefficient in a computer simulation with two to five predictors, moderate intercorrelation, large sample size (1000), and small to moderate proportion of missing values (1%-10%). When varying the proportion of missing values and the number of

predictors in a computer simulation, Gleason and Staelin (1975) found pairwise deletion better than mean substitution in reproducing the covariance matrix when the average intercorrelation was 0.25 or greater and the sample size was small. When the sample size was large, pairwise deletion performed better with an average intercorrelation of 0.15. Buck (1960), Kim and Curry (1977), and Gleason and Staelin (1975), all used random deletion of data to create missing values.

Average intercorrelation and number of variables are the two factors that produce the largest differences in reproducing the covariance matrix. Keeping sample size and proportion of missing values constant, when the average intercorrelation is low and the number of variables small, listwise deletion performs better. If the average intercorrelation is above 0.2 and the number of variables larger than three, pairwise deletion performs better (Buck, 1960; Kim and Curry, 1977; Gleason & Staelin, 1975).

Estimation of Missing Values

Missing values are estimated and imputed to avoid non-representation resulting from dropping cases; to avoid power loss; to capitalize on inherent information in the missing/nonmissing pattern; and to utilize information in the other variables (Cohen & Cohen, 1983). The estimation procedure assumes that the values are missing at random and

that the proportion of missing is not excessive. Multivariate methods assume in addition that each missing variable is highly correlated with one or more other variables (Frane, 1976).

Mean Substitution

Mean substitution, attributed to Wilks' (1932) fills in a variable's missing values by the mean of that variable. Assuming a normal distribution, the sample mean of that variable is the optimal estimate of its most probable value. If the distribution is not normal, the median for the variable is substituted. This method does not alter the sample mean, but artificially reduces the variance for the treated variable. This results in a reduction in the levels of association between the variables which introduces error in the explanatory variables and may bias regression slopes (Anderson et al. 1983; Gleason & Staelin, 1975; Hertel, 1976). It further concentrates all the imputed values at the mean thus distorting and creating spikes in the distribution (Kalton & Kish, 1981).

In a computer simulation, Afifi and Elashoff (1967) discovered that mean substitution was better than listwise deletion in estimating the regression coefficient with low average intercorrelation (less than 0.3), small samples (less than 70), two variables, and a high proportion of missing values. This study did not vary the number of

variables. Gleason and Staelin (1975) found mean substitution better than pairwise deletion in estimating the covariance matrix with low average intercorrelations (less than 0.2) and small sample sizes (less than 50). Timm (1970) determined that mean substitution was better than Buck's procedure and listwise deletion in estimating the covariance matrix when the variables have low average intercorrelations (less than 0.3) and the proportion of missing values is high (20%). Chan and Dunn (1972) in a simulation study found mean substitution better than listwise deletion, pairwise deletion, and two regression methods in discriminating between groups when the determinant of the sample correlation matrix was large, the sample size was small (15-35), and the proportion of missing values was high (20%).

In contrast to listwise and pairwise deletion, mean substitution performs better with small sample sizes and a high proportion missing. The estimate of a variable's missing values produced by mean substitution is dependent only on that variable's known values. Therefore, it is not influenced by the number of variables nor by the pattern of missing values. With low average intercorrelations, it is similar to the listwise deletion method. The major problem with this method was an artificial reduction in the variance of treated variables which reduced the correlation.

Regression Methods

Regression as an imputation method has many variations. The variations rely on information from other variables to estimate missing values. As the average intercorrelation and the number of variables from which these methods can obtain information increases, the regression methods, theoretically, perform better. Too many variables, however, can cause problems with overprediction (Kaiser & Tracy, 1988) and too high an average intercorrelation can result in a singular matrix. In these cases, regression does not perform well.

Variations of the regression method include differences in methods of developing the initial correlation matrix (listwise deletion, pairwise deletion, and mean substitution) and the presence or absence of iteration procedures.

Buck's Procedure

Buck (1960) was the first to use regression to estimate missing values. In this procedure each variable with a missing value is regressed on the non-missing variables. The initial correlation matrix used to develop the regression equation for each missing value is produced by listwise deletion. The resulting equations are used to provide estimates of the missing values. These estimates are then inserted into the incomplete data set.

Buck's procedure provides a consistent estimate of variable means conditioned on variables that are used as predictors in the regression equation. Since it projects incomplete cases to the regression line, the variance and covariance of treated variables are underestimated (Little & Rubin, 1987). It also assumes the regression is linear which becomes a tenuous assumption if the imputed values are extrapolated beyond the range of the data.

Using real data, Buck (1960) found this regression variation superior to listwise deletion in estimating variable means and standard deviations with low average intercorrelation, four variables, 12% missing values, and a sample size of 72. Chan and Dunn (1972) found this method superior to listwise deletion, pairwise deletion, mean substitution, and another regression variation in estimating the linear discriminant function when the determinant of the correlation matrix was small, proportion of missing values was high (20%), and there were 2 to 8 variables.

Timm (1970) found Buck's procedure superior to listwise deletion and mean substitution in estimating the covariance matrix when the average intercorrelation is low (0.2) and either the number of variables was greater than two with 1% missing values, or when the number of variables was greater than five and the proportion of missing values was moderate (10%) to high (20%). When the average intercorrelation was

high, Buck's procedure was better with moderate to high proportion of missing values and more than two variables. Sample size did not affect these results.

Buck's procedure uses listwise deletion for the initial correlation matrix, it becomes less effective than some other regression variations as the number of variables increases. Chan et al. (1976) abandoned using Buck's procedure for eight variables because the initial correlation matrix was based on a very small set of complete observations.

Complete Data Methods

Gleason and Staelin (1975) criticized Buck's procedure for failing to use all the sample information in the initial correlation matrix. They recommended modification of the regression technique by beginning with a complete data correlation matrix either by pairwise deletion or mean substitution. This variation enables the maximum use of available information to develop regression equations. This method was found superior to mean substitution whenever the average intercorrelation was greater than 0.2.

Chan et al. (1976) found the regression variation using mean substitution in the initial correlation matrix superior to listwise deletion, pairwise deletion, mean substitution, and Buck's procedure in discriminating between groups with large determinants. It was also better with medium

determinants having four or more variables. This method was recommended unless a near-singular correlation matrix is expected.

Iterative Regression

Iterative regression begins with an initial correlation matrix that is produced by the listwise deletion (Buck's procedure), pairwise deletion, or mean substitution methods. Estimates for the missing values are obtained and imputed as in the prior regression procedures. Then each variable for which there was a missing value (now replaced by an estimate) is regressed on all the other variables. A new estimate is obtained and replaces the initial one. The development of new regression equations and estimates is continued until subsequent iterations do not produce a significant change in the estimates or when the maximum number of iterations are completed (Beale & Little, 1975).

Dempster, Laird, and Rubin (1977) recommended the use of the EM (expectation maximization) algorithm which imputes estimates simultaneously in an iterative procedure. In the discussion following this paper, Healy (Dempster et al., 1977) declared the EM algorithm equivalent to the Jacobi method of solving simultaneous equations. The alternative is to estimate values and to adjust them one at a time using the Gauss-Seidel method. Both methods converge to the same

final estimates, but the speed of convergence differs. The EM algorithm was advocated to hasten convergence.

Beale and Little (1975) introduced the 'Iterated Buck' method which produces a correlation matrix using listwise deletion. Once initial estimates of the missing values are obtained, iterative regression is performed. This modified maximum likelihood method does not assume multivariate normality as do true maximum likelihood methods. Using computer simulation, they found this method superior to listwise deletion in all sample sizes with 5% to 40% missing values and three to five variables. It was also superior to Buck's procedure except for three cases. Most of the iterations had converged within 10 runs, but one regression required 171 iterations. It was recommended to set the maximum number of iterations prior to beginning this iterative process.

Using real data, Jackson (1968) compared the mean substitution method with iterative regression in which initial correlation matrix was determined by mean substitution. Neither the average intercorrelation nor the proportion of missing values were reported in this study. Twenty-seven variables were used in discriminating groups. Methods were judged based on the percentage of instances in which each method correctly classified the subjects. The iterative regression procedure was found marginally better

than the mean substitution method. She limited the number of iterations to six and recommended that future researchers use a higher maximum of iterations.

Although any estimate used for missing values would enable the use of all available data, Little and Rubin (1987) cautioned that the use of pairwise deletion in the initial correlation matrix may yield a singular covariance matrix. If it occurs, iterative regression will be a problem. Mean substitution to produce the initial correlation matrix may, however, speed the convergence of iterative procedures (Dempster et al, 1977).

Factors Influencing Missing Data Methods

Sample size

Because listwise deletion relies on the number of complete cases in the sample, larger sample sizes improve performance if the proportion of missing values does not increase. As the sample size increases, the stability of the correlation matrix increases. Therefore, pairwise deletion and the regression methods also benefit from increased sample size.

Compared to other methods, mean substitution appears to estimate the regression coefficient (Afifi & Elashoff, 1967) and the covariance matrix (Timm, 1970) better with small samples. It is unclear whether mean substitution

performed better due to small samples, the other methods performed relatively worse, or the proportion of missing values influenced these results.

Number of Variables

As the number of variables increases, methods relying on information in other variables, such as regression, perform better. Listwise deletion deteriorates due to elimination of cases. Although Buck's procedure is a regression method, listwise deletion in the initial correlation matrix may reduce the number of cases producing this matrix to an unusable number (Chan et al., 1976).

Proportion of Missing Values

As the proportion of missing values increases, all methods deteriorate. Because listwise deletion relies on the number of complete cases in the sample, a smaller proportion of incomplete cases improves performance. If values are missing randomly, an increase in missing values will result in an increase in incomplete cases. Therefore, an increase in missing values will adversely affect the performance of listwise deletion. Methods relying on listwise deletion in the initial correlation matrix are also affected.

Average Intercorrelation

As the average intercorrelation increases, those methods capitalizing on information in other variables

(regression and pairwise deletion) increase in effectiveness.

Summary

Listwise deletion has been shown to be more effective with large samples, low average intercorrelation, less than four variables and a small proportion of missing values. Problems for a researcher using this method include a reduction in power and an increase in standard error due to reduced sample size and the elimination of sub-populations.

Pairwise deletion is more effective with large samples, higher average intercorrelation, and a small proportion of missing values. The number of variables does not influence results. The primary problem encountered when using this method is production of a non-symmetric covariance matrix.

Mean substitution is more effective with small samples, low average intercorrelation, and a high proportion of missing values. This method is not influenced by the number of variables. The problem with this method is an artificial reduction in the variance of treated variables.

The regression methods rely on information contained in non-missing values of other variables to provide estimates of missing values. Theoretically, the more variables considered that provide additional information, the better the estimate. Too many variables however can result in overprediction.

Buck's procedure is most effective with higher average intercorrelation, small proportion of missing values, and between two and eight variables. Sample size has shown no effect in the studies cited. Since larger samples produce a more stable correlation matrix, large sample size should be beneficial to this method.

By using mean substitution or pairwise deletion in computing the initial correlation matrix, all values are considered for regression estimates. This should provide estimates that are closer to the real values. Mean regression is more effective with more than four variables and higher average intercorrelation. Because this method uses all the sample values, the stability of the initial correlation matrix is maximized.

Iterative regression is a repetitive estimation until further estimates do not change. This process can be very slow. Since the first iteration produces the largest change the difference in the outcome of 1-iteration and iterative regression may not be large.

CHAPTER III

Method

This chapter consists of sections describing the selection of variables, construction of test samples, methods of handling missing data, and analysis. The effectiveness of listwise deletion, pairwise deletion, mean substitution, Buck's procedure, mean regression, 1-iteration regression, and iterative regression in reproducing the true covariance matrix and variable means was compared in a factorial design.

Factors that were manipulated are: two levels of number of variables (4 and 8), two levels of average intercorrelation (0.2 and 0.4), two levels of sample size (200 and 2000), and three levels of proportion of incomplete cases (10%, 20%, and 40%). A brief description of the methodology used follows.

Variables for this study were chosen from those related to parental involvement and met the number of variables and average intercorrelation requirements. To compare these methods empirically, cases with no missing values were selected and placed in a non-missing population. A target sample of the desired size was randomly selected from this population. The covariance matrix and variable means of the target sample were recorded. Incomplete cases formed the missing population. A random sample of the proportion of

incomplete cases was then selected. Each of these cases was matched to a case in the target sample representing non-missing cases. Whenever a best match was found, the case with missing values replaced the match in the target sample. The resulting sample was called the test sample. Each of the 24 test samples thus constructed had their missing values treated by the seven methods of handling missing data. The relative effectiveness of each method was determined by comparing the resulting covariance matrix and variable means with those of the target sample.

Selection of Variables

Variables for this study were selected from the student and the parent supplements of the National Education Longitudinal Study of 1988 (NELS-88). NELS-88 was designed as a nationally representative two-stage stratified probability sample with schools selected in stage one and students at the second stage. The final student sample included 24,599 students (Ingels et al., 1990a).

The student supplement includes responses from a 5-minute questionnaire and a series of achievement test scores. Parents of participating students completed a 30-minute questionnaire forming the parental supplement (Ingels et al., 1990b). Over 100 variables were chosen from these two supplements for consideration either in forming composite variables, or as contributors to a latent

construct in the various models of a study concerning the effects of parental involvement in middle schools (Keith et al., 1992). The variables used in this study were selected from those in the parental involvement study.

The selection of variables involved decisions about the number of variables and their average intercorrelation. The number of variables, 4 or 8, were chosen arbitrarily. However, the selection of specific variables depended on how they intercorrelated with each other. Two levels of average intercorrelation, as determined by Kaiser's Gamma, were used. The levels were 0.2 (range = 0.103 to 0.270) and 0.4 (range = 0.265 to 0.722).

The variables for both four-variable groups were chosen from the parental supplement of NELS-88. All four variables in these two samples were measured on a scale of one to four, with one being none and four representing more than five. The first two variables in the 0.4 average intercorrelation group (X1 - BYP57A and X2 - BYP57B) were school initiated contact with the parents concerning the academic performance or the academic program of the student. The last two variables in this sample (X3 - BYP58A and X4 - BYP58B) were parent initiated contact with the school concerning the academic performance or the academic program of the student and are shown below.

| Name | NELS Code | Description |
|------|-----------|---|
| X1 | BYP57A | CONTACTED ABOUT ACADEMIC PERFORMANCE |
| X2 | BYP57B | CONTACTED ABOUT ACADEMIC PROGRAM |
| X3 | BYP58A | CONTACTED SCHOOL ABOUT ACADEMIC PERFORM |
| X4 | BYP58B | CONTACTED SCHOOL ABOUT ACADEMIC PROGRAM |

The first two variables in the 0.2 average intercorrelation sample (X1 - BYP57C and X2 - BYP57F) were school initiated contact with the parents concerning the student's high school course selection and about school fund raising. The last two variables in this sample (X3 - BYP58A and X4 - BYP58E) were parent initiated contact with the school concerning the student's academic performance and information for school records and are shown below.

| Name | NELS Code | Description |
|------|-----------|---|
| X1 | BYP57C | CONTACTED ABOUT H.S. COURSE SELECTION |
| X2 | BYP57F | CONTACTED ABOUT SCHOOL FUND RAISING |
| X3 | BYP58A | CONTACTED SCHL ABOUT ACADEMIC PERFORMAN |
| X4 | BYP58E | CONTACTED SCHL ABOUT INFO FOR SCH RECOR |

The first two variables in each sample were selected from responses to the subdivisions of question 57. The last two variables in each sample were selected from responses to the subdivisions of question 58. The distinguishing characteristic between the first two variables and the last two in each sample is who initiated contact (parent or school). These variables were chosen because they are from the same categories but still exhibit the desired average intercorrelation. Table 1 contains the correlation matrices and descriptive statistics for these samples.

Table 1

Correlation Matrices and Descriptive Statistics of Four Variables

| Correlation 0.2 | | | | |
|------------------|------|------|------|------|
| | X1 | X2 | X3 | X4 |
| X1 | 1.00 | | | |
| X2 | .181 | 1.00 | | |
| X3 | .125 | .154 | 1.00 | |
| X4 | .114 | .180 | .243 | 1.00 |
| Mean | 1.42 | 1.64 | 1.76 | 1.42 |
| SD ^a | .55 | .87 | .87 | .56 |
| Freq | 170 | 148 | 735 | 741 |
| PMP ^b | .166 | .144 | .717 | .723 |
| PMN ^c | .007 | .006 | .031 | .031 |
| Correlation 0.4 | | | | |
| | X1 | X2 | X3 | X4 |
| X1 | 1.00 | | | |
| X2 | .535 | 1.00 | | |
| X3 | .451 | .282 | 1.00 | |
| X4 | .303 | .430 | .595 | 1.00 |
| Mean | 1.89 | 1.47 | 1.76 | 1.44 |
| SD ^a | .97 | .70 | .87 | .67 |
| Freq | 154 | 252 | 740 | 828 |
| PMP ^b | .137 | .224 | .657 | .735 |
| PMN ^c | .006 | .011 | .031 | .035 |

Note. Freq = Frequency of occurrence after adjustment for more than half missing values or legitimate skips.
^aStandard deviation. ^bProportion missing values per missing population. ^cProportion missing values/NELS-88.

Both eight variable samples were selected from the student supplement of NELS-88. The first four variables in the 0.4 average intercorrelation sample (X1 - BYS81A, X2 - BYS81B, X3 - BYS81C, and X4 - BYS81D) were student report of English, math, science, and social studies grades from grade 6 until now (grade 8). The last four variables in this sample (X5 - BYTXHSTD, X6 - BYTXRSTD, X7 - BYTXMSTD, and X8 - BYTXSSTD) were standardized test scores of the student in history/citizenship/geography, reading, mathematics, and science. The first four variables in this group were scaled from one to six. All variables in this sample are reproduced below.

| Name | NELS Code | Description |
|------|-----------|---------------------------------------|
| X1 | BYS81A | ENGLISH GRADES FROM GRADE 6 UNTIL NOW |
| X2 | BYS81B | MATH GRADES FROM GRADE 6 UNTIL NOW |
| X3 | BYS81C | SCIENCE GRADES FROM GRADE 6 UNTIL NOW |
| X4 | BYS81D | SOC. STUDIES GRDS FRM GRADE 6 TIL NOW |
| X5 | BYTXHSTD | HISTORY/CIT/GEOG STANDARDIZED SCORE |
| X6 | BYTXRSTD | READING STANDARDIZED SCORE |
| X7 | BYTXMSTD | MATHEMATICS STANDARDIZED SCORE |
| X8 | BYTXSSTD | SCIENCE STANDARDIZED SCORE |

The eight variables in the 0.2 average intercorrelation sample were selected from various groups of questions including whether other students in class view the student in question as a good student (X1 - BYS56C), the student's ability group for mathematics (X2 - BYS60A), whether the student discusses things studied in class with parents (X3 - BYS36C), how far in school the student's mother wants the student to go (X4 - BYS48B), how often the

student talked to father about planning the high school program (X5 - BY550A), how much time the student spends on math homework each week (X6 - BY579A), how often the student comes to class without homework (X7 - BY578C), and the student's reading standardized score (X8 - BYTXRSDT). The first seven variables in this group were scaled items. These variables are reproduced below.

| Name | NELS Code | Description |
|------|-----------|---|
| X1 | BY556C | STUDENTS IN CLASS SEE R AS GOOD STUDENT |
| X2 | BY560A | R'S ABILITY GROUP FOR MATHEMATICS |
| X3 | BY536C | DISCUS THNGS STUDIED IN CLSS WITH PRNTS |
| X4 | BY548B | HOW FAR IN SCHL R'S MOTHER WNTS R TO GO |
| X5 | BY550A | TALK TO FATHER ABOUT PLANNING H.S. PROG |
| X6 | BY579A | TIME SPENT ON MATH HOMEWORK EACH WEEK |
| X7 | BY578C | HOW OFTEN COME TO CLASS WITHOUT HOMEWRK |
| X8 | BYTXRSTD | HIST/CIT/GEOG STANDARDIZED SCORE |

Table 2 contains the correlation matrices and descriptive statistics for these samples.

Creating a Test Sample

Cases in the NELS-88 data set that did not have a missing value on the selected variable were separated and were named non-missing population. Cases with one or more missing values were called missing population. Any case that contained missing values for over half of the selected variables or contained a legitimate skip was not included in the missing population. This deletion, when it happened, changed the proportion of incomplete cases and is shown in Appendix A (Tables 1-4).

Table 2

Correlation Matrices and Descriptive Statistics of Eight Variables

| Correlation 0.2 | | | | | | | | |
|------------------|-------|-------|------|------|-------|-------|-------|-------|
| | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| X1 | 1.00 | | | | | | | |
| X2 | -.199 | 1.00 | | | | | | |
| X3 | .170 | -.103 | 1.00 | | | | | |
| X4 | .158 | -.156 | .154 | 1.00 | | | | |
| X5 | .165 | -.131 | .266 | .170 | 1.00 | | | |
| X6 | .149 | -.157 | .185 | .152 | .150 | 1.00 | | |
| X7 | .270 | -.109 | .151 | .115 | .146 | .142 | 1.00 | |
| X8 | .228 | -.230 | .224 | .291 | .152 | .243 | .153 | 1.00 |
| Mean | 1.72 | 2.22 | 2.42 | 5.04 | 1.09 | 1.99 | 3.00 | 50.50 |
| SD ^a | .61 | 1.19 | .69 | 1.19 | .76 | 1.48 | .86 | 10.08 |
| Freq | 644 | 513 | 315 | 1469 | 600 | 969 | 1310 | 870 |
| PMP ^b | .140 | .111 | .068 | .318 | .130 | .210 | .284 | .189 |
| PMN ^c | .026 | .021 | .013 | .060 | .025 | .040 | .054 | .036 |
| Correlation 0.4 | | | | | | | | |
| | X1 | X2 | X3 | X4 | X5 | X6 | X7 | X8 |
| X1 | 1.00 | | | | | | | |
| X2 | .383 | 1.00 | | | | | | |
| X3 | .464 | .397 | 1.00 | | | | | |
| X4 | .476 | .364 | .539 | 1.00 | | | | |
| X5 | .320 | .265 | .394 | .433 | 1.00 | | | |
| X6 | .371 | .276 | .399 | .429 | .722 | 1.00 | | |
| X7 | .367 | .399 | .422 | .414 | .691 | .710 | 1.00 | |
| X8 | .296 | .275 | .393 | .373 | .712 | .701 | .717 | 1.00 |
| Mean | 2.06 | 2.09 | 2.22 | 2.22 | 50.51 | 50.50 | 50.64 | 50.40 |
| SD ^a | .99 | 1.04 | 1.11 | 1.16 | 10.06 | 10.08 | 10.22 | 10.13 |
| Freq | 291 | 269 | 301 | 381 | 900 | 894 | 915 | 1005 |
| PMP ^b | .180 | .166 | .186 | .236 | .557 | .553 | .566 | .622 |
| PMN ^c | .012 | .011 | .012 | .016 | .037 | .036 | .037 | .041 |

Note. Freq = Frequency of occurrence after adjustment for more than half missing values or legitimate skips.

^aStandard deviation. ^bProportion missing values per missing population. ^cProportion missing values/NELS-88.

Target samples of 200 and 2000 cases were randomly selected from the population of non-missing cases. The levels of sample size were chosen to represent small and large sized samples.

The method used by Little and Su (1987) was followed in matching cases from the missing sample to those in the target sample using the non-missing variables. The purpose was to preserve the pattern of missing values in the target sample, and to prevent changes in the values of the non-missing values that might affect the performance of the missing value methods. A random sample of incomplete cases (missing sample) from the population of incomplete cases was selected. The values of the non-missing variables in the incomplete case were compared to the values of the corresponding variables of the target sample. Matching took place sequentially from the first record in the target sample and continued until a match was found. Whenever the values matched, the target sample case was removed and replaced by the missing sample case. When values did not match, the missing case was compared to the second case in the target sample. This process continued until an exact match was found. When an exact match was not found, the closest match was used.

Criteria for determining a closest match was based on the zero-order correlation of the variable containing a

missing value with the other variables. The variable with the highest zero-order correlation with the one containing a missing value was matched first, followed by the variable with the next highest zero-order correlation. This process established the order in which values on the incomplete case were to be matched with values on the complete cases (target sample). The complete case that had the closest match was replaced by the incomplete case from the missing sample. When multiple cases from the target sample had an exact match with the observed values of the incomplete case, the complete case was selected at random for replacement. This procedure was repeated for all the cases in the missing sample.

The proportion of missing values was also computed to assist comparison with simulation studies cited in the literature review and is reported in Table 3.

Methods of Handling Missing Data

Seven methods of handling missing data were chosen for comparison in this study: listwise deletion, pairwise deletion, mean substitution, Buck's procedure, mean regression, 1-iteration regression, and iterative regression. The first three methods were chosen because they represented the methods of choice for most computer users. Listwise deletion is the default for many computer programs. Buck's procedure was chosen because it does not

Table 3

Comparison of Proportion of Incomplete Cases to Proportion of Missing Values

| Incomplete Cases | Proportion Missing | |
|------------------|---------------------|---------------------|
| | Values ^a | Values ^b |
| Correlation 0.2 | | |
| Four Variables | | |
| 10% | 4% | 4% |
| 20% | 9% | 9% |
| 40% | 17% | 18% |
| Eight Variables | | |
| 10% | 2% | 2% |
| 20% | 4% | 4% |
| 40% | 7% | 7% |
| Correlation 0.4 | | |
| Four Variables | | |
| 10% | 4% | 4% |
| 20% | 9% | 9% |
| 40% | 16% | 17% |
| Eight Variables | | |
| 10% | 4% | 4% |
| 20% | 8% | 8% |
| 40% | 16% | 15% |

Note. ^aSmall sample. ^bLarge sample.

use mean substitution in the initial correlation matrix. Iterative regression was selected as a representative of modified maximum likelihood methods. Mean regression and 1-iteration regression are initial steps of the iterative regression procedure. Compared to the other methods, the regression methods are less preferred because they are more demanding in computer time and are difficult to implement.

Listwise deletion deletes a case from the sample if it has one or more variable values missing. Of the methods used, it will cause the greatest loss of data.

In pairwise deletion, all pairs having both observations are used to compute covariances. Means and variances are computed on all available observations. This method may yield a non-symmetric covariance matrix.

The mean substitution method substitutes missing values with their variable mean. This method artificially reduces the variance for the treated variable and consequently affects correlation with the other variables.

According to Buck's procedure, each variable with missing values is treated as a criterion variable and regressed on all other non-missing variables. Listwise deletion is used to produce the correlation matrix and to develop regression equations. The resulting equations are used to estimate the missing values. The obtained estimates

replace the missing value code for each case yielding a complete data matrix.

In mean regression, missing values are replaced by their variable means to produce the initial correlation matrix. The regression equation is developed by regressing the variable having a missing value on all the remaining variables. The estimate obtained replaces the variable mean used as an initial estimate. This method allows all predictors to contribute to the missing value estimate. In spite of the concerns that the variable mean used as an initial estimate of missing values will reduce the variable's variance and would affect its correlation with the other variables, this procedure is recommended by Dempster et al. (1977).

One-iteration regression is an extension of the mean regression method. Each variable for which there was a missing value (now replaced by the mean regression method estimate) is regressed on all the other variables. New estimates are obtained and replace the prior ones. This method was suggested by Little and Su (1983) as a compromise solution for iterative regression.

Iterative regression continues the one-iteration method until subsequent iterations do not produce a significant change in the variable means and standard deviations, or when the number of iterations exceed 10.

Analysis

The procedures outlined in this chapter were implemented using SPSS/PC+ software (Norusis, 1990). Analysis was accomplished using SPSS/PC+ (Norusis 1990; 1988a), LISREL 7 (Joreskog & Sorbom, 1989), and Lotus 3.1 (LeBlond & Cobb, 1990).

Multi-sample analysis in LISREL (Joreskog & Sorbom, 1989, chap. 9) was used to test the equality of the sample covariance matrix produced by various missing data handling methods to the covariance matrix of the target sample. Within multi-sample analysis, models are compared by goodness of fit functions. In determining the fit of each sample, LISREL iteratively computes the population covariance matrix, $\underline{\Sigma}$, until the best fit between $\underline{\Sigma}$ and the sample covariance matrix, \underline{S} , is obtained. Each \underline{S} is then tested for fit as if it were a sampling fluctuation of $\underline{\Sigma}$ (Hayduk, 1989). The fit of each model is estimated without constraint. Measures of goodness of fit for each group are given. Then parameters which are specified to be equal across groups are replaced by their mean values. Chi-square (χ^2) measures the fit of all models in all groups including constraints to the data from all groups (Joreskog & Sorbom, 1989).

To compare the covariance matrices in this study, two runs were required for each test sample. In the first run,

all the elements of both covariance matrices were free to vary. This tested the hypothesis that both samples were selected from the same population. In the second run, the elements of the covariance matrix of the target sample were free to vary, but the elements of the test sample covariance matrix were constrained to be equal to those of the first. Each run produced a χ^2 measure of fit. The difference between these χ^2 s is distributed as χ^2 . The resultant χ^2 "provides a test for the significance of the differences between the covariance matrices" (Hayduk, 1989). Tatsuoka (1988) says this compares the fit of two models; one which frees the dispersions to vary, and one which constrains the dispersions to be equal.

The variable means produced by each method were compared with the corresponding mean values of the target sample using the MANOVA subroutine in SPSS/PC+ for every method except pairwise deletion. Since the MANOVA subroutine does not accept pairwise deletion, the vector of variable means produced by pairwise deletion was compared to that of the target sample using Lotus 3.1. One advantage of using MANOVA is that the univariate results are produced automatically after a multivariate test.

Chapter IV

Results

When there were four variables and small samples (Table 4), all methods effectively ($p > .05$) reproduced the target sample covariance matrix when the proportion of incomplete cases was 20% or less. When, however, the proportion of incomplete cases increased to 40%, only listwise deletion was effective ($p > .05$) when the average intercorrelation was 0.2. Listwise deletion, pairwise deletion, and mean substitution were effective ($p > .05$) when the average intercorrelation was 0.4.

The listwise and pairwise deletion methods effectively ($p > .05$) reproduced the four-variable target sample covariance matrix in large samples at both levels of average intercorrelation and across all proportions of incomplete cases. The remaining methods adequately ($p > .05$) reproduced the target sample covariance matrix only when the proportion of incomplete cases was 10% and the average intercorrelation was 0.4. As the proportion of incomplete cases increased, methods other than listwise and pairwise deletion were less effective ($p < .05$) progressing from a probability of reproducing the target sample covariance matrix of 0.05 to a probability of 0.01.

Table 4

Results of Retaining Four-Variable Covariance Matrix

| Method | Proportion Incomplete | | |
|---|-----------------------|---------|----------|
| | 10% | 20% | 40% |
| <u>Small samples, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | 1.03 | 2.80 | 17.93 |
| Pairwise Deletion | .81 | 1.93 | 19.27* |
| Mean Substitution | 2.25 | 6.88 | 52.44** |
| Bucks Procedure | 2.39 | 7.59 | 66.25** |
| Mean Regression | 2.34 | 7.27 | 62.20** |
| 1 Iteration | 2.39 | 7.58 | 65.97** |
| Iterative Regression | 2.39 | 7.76 | 68.16** |
| <u>Small samples, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .53 | 2.30 | 4.96 |
| Pairwise Deletion | .64 | 2.67 | 7.18 |
| Mean Substitution | .71 | 10.39 | 14.58 |
| Bucks Procedure | 1.08 | 10.45 | 21.30* |
| Mean Regression | .76 | 10.51 | 19.37* |
| 1 Iteration | .89 | 10.25 | 21.31* |
| Iterative Regression | 1.10 | 10.15 | 22.94* |
| <u>Large samples, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | 10.33 | 11.80 | 5.12 |
| Pairwise Deletion | 8.16 | 5.21 | 4.61 |
| Mean Substitution | 22.56* | 39.54** | 121.57** |
| Bucks Procedure | 23.99* | 23.81* | 119.20** |
| Mean Regression | 23.55* | 45.72** | 121.57** |
| 1 Iteration | 23.99* | 46.80** | 121.01** |
| Iterative Regression | 23.99* | 46.12** | 121.27** |
| <u>Large samples, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .76 | 1.88 | 12.57 |
| Pairwise Deletion | .70 | 1.08 | 4.28 |
| Mean Substitution | 5.12 | 28.36** | 116.58** |
| Bucks Procedure | 7.27 | 32.01** | 177.06** |
| Mean Regression | 6.96 | 31.05** | 162.85** |
| 1 Iteration | 7.19 | 31.47** | 183.77** |
| Iterative Regression | 7.27 | 31.24** | 198.45** |

Note. γ = Kaiser's Gamma. df=10. * $p < .05$. ** $p < .01$.

With four variables in small samples (Table 4), all methods adequately ($p > .05$) reproduced the target sample covariance matrix as the average intercorrelation increased from 0.2 to 0.4. When the sample size was large and 10% of the cases were incomplete, mean substitution, and the regression methods adequately ($p > .05$) reproduced the target sample covariance matrix when the average intercorrelation was 0.4, but did not when the average intercorrelation was 0.2. When the sample size was large with 20% incomplete cases, Buck's procedure was less effective when the average intercorrelation was 0.4 ($p < .01$) than when the average intercorrelation was 0.2 ($p < .05$).

As the sample size increased from small to large samples with four variables (Table 4), fewer of the methods of handling missing data could adequately ($p > .05$) reproduce the target sample covariance matrix. When the average intercorrelation was 0.2, mean substitution, and all the regression methods progressed from effectively ($p > .05$) reproducing the target sample covariance matrix in small samples to ineffective ($p < .05$) with large samples. On the other hand, pairwise deletion was ineffective ($p < .05$) with 40% incomplete cases and small samples, but was effective ($p > .05$) under the same conditions with large samples. When the average intercorrelation was 0.4, mean substitution and the regression methods progressed from effective ($p > .05$)

with 20% incomplete cases or ineffective ($p < .05$) with 40% incomplete cases with small samples to ineffective ($p < .01$) with large samples.

When the sample size was small with eight variables, (Table 5) all methods adequately reproduced the target sample covariance matrix under all conditions except iterative regression. Iterative regression was ineffective ($p < .05$) with 40% incomplete cases and an average intercorrelation of 0.4. When the sample size increased from small to large samples, listwise deletion was the only method that did not adequately ($p < .01$) reproduce the target covariance matrix when the average intercorrelation was 0.2. When the average intercorrelation increased to 0.4 only listwise and pairwise deletion adequately ($p > .05$) reproduced the target sample covariance matrix.

When the number of variables increased from four (Table 4) to eight (Table 5) with small samples, methods that failed to reproduce ($p < .05$) the target sample covariance matrix with four variables were effective ($p > .05$) with eight variables, except for iterative regression. When the sample size was large, methods that were ineffective ($p < .05$) with four variables at 10% and 20% incomplete cases were effective ($p > .05$) with eight variables. When 40% of the cases were incomplete and the average intercorrelation was 0.2, listwise deletion adequately reproduced the target

Table 5

Results of Retaining Eight-Variable Covariance Matrix

| Method | Proportion Incomplete | | |
|---|-----------------------|-------|----------|
| | 10% | 20% | 40% |
| <u>Small samples, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | 2.18 | 4.83 | 13.21 |
| Pairwise Deletion | 2.42 | 3.87 | 8.86 |
| Mean Substitution | 3.01 | 4.19 | 11.91 |
| Bucks Procedure | 2.97 | 4.04 | 14.77 |
| Mean Regression | 3.01 | 4.26 | 16.42 |
| 1 Iteration | 3.01 | 4.42 | 17.80 |
| Iterative Regression | 3.01 | 4.46 | 18.49 |
| <u>Small samples, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | 2.46 | 5.00 | 18.57 |
| Pairwise Deletion | 3.36 | 10.74 | 22.71 |
| Mean Substitution | 4.07 | 11.67 | 35.87 |
| Bucks Procedure | 2.57 | 8.53 | 44.49 |
| Mean Regression | 4.20 | 9.42 | 45.66 |
| 1 Iteration | 4.16 | 8.92 | 48.56 |
| Iterative Regression | 4.17 | 8.87 | 50.86* |
| <u>Large samples, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | 4.84 | 14.34 | 71.54** |
| Pairwise Deletion | 4.59 | 8.80 | 15.91 |
| Mean Substitution | 4.95 | 12.70 | 35.99 |
| Bucks Procedure | 5.86 | 11.59 | 41.66 |
| Mean Regression | 5.86 | 12.55 | 43.33 |
| 1 Iteration | 6.03 | 12.38 | 46.24 |
| Iterative Regression | 6.03 | 12.38 | 47.28 |
| <u>Large samples, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | 5.83 | 7.44 | 49.28 |
| Pairwise Deletion | 5.84 | 9.88 | 20.21 |
| Mean Substitution | 12.63 | 26.82 | 106.40** |
| Bucks Procedure | 17.10 | 29.40 | 160.20** |
| Mean Regression | 14.97 | 27.18 | 152.23** |
| 1 Iteration | 15.91 | 27.21 | 159.70** |
| Iterative Regression | 17.23 | 29.02 | 171.42** |

Note. γ = Kaiser's Gamma. df=36. * $p < .05$. ** $p < .01$.

sample covariance matrix with four variables, but did not with eight variables. On the other hand, Mean substitution and the regression methods were ineffective ($p < .01$) with four variables but adequately reproduced the target covariance matrix with eight variables when the average intercorrelation was low. When the average intercorrelation was 0.4, methods that were ineffective ($p < .01$) with four variables remained ineffective ($p < .01$) with eight variables.

When there were four variables (Table 6), all methods adequately ($p > .05$) reproduced the target sample variable means over all conditions. When there were eight variables (Table 7), all methods adequately ($p > .05$) reproduced the target sample variable means when the sample size was small. When the sample size was large with eight variables, however, listwise deletion was ineffective with 20% and 40% incomplete cases ($p < .05$) when the average intercorrelation was 0.2. All other methods adequately reproduced the target sample variable means under these conditions. When the average intercorrelation increased to 0.4 all methods adequately reproduced the target sample variable means with 20% incomplete cases or less. With 40% incomplete cases, only pairwise deletion adequately reproduced the target mean vector.

As the proportion of incomplete cases increased, the effectiveness of methods in reproducing the target sample

Table 6

Results of Four-Variable Means Comparison

| Method | Proportion Incomplete | | |
|---|-----------------------|------|-------|
| | 10% | 20% | 40% |
| <u>Small samples^a, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | .238 | .049 | .905 |
| Pairwise Deletion | .291 | .061 | .947 |
| Mean Substitution | .151 | .076 | 1.488 |
| Bucks Procedure | .163 | .065 | 1.232 |
| Mean Regression | .161 | .071 | 1.400 |
| 1 Iteration | .162 | .069 | 1.326 |
| Iterative Regression | .162 | .069 | 1.232 |
| <u>Small samples^a, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .004 | .240 | .580 |
| Pairwise Deletion | .029 | .034 | .253 |
| Mean Substitution | .031 | .040 | .386 |
| Bucks Procedure | .011 | .050 | .056 |
| Mean Regression | .011 | .023 | .309 |
| 1 Iteration | .011 | .031 | .210 |
| Iterative Regression | .011 | .051 | .108 |
| <u>Large samples^b, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | .557 | .402 | .515 |
| Pairwise Deletion | .380 | .173 | .123 |
| Mean Substitution | .421 | .215 | .198 |
| Bucks Procedure | .394 | .387 | .263 |
| Mean Regression | .393 | .197 | .196 |
| 1 Iteration | .389 | .197 | .188 |
| Iterative Regression | .388 | .197 | .175 |
| <u>Large samples^b, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .030 | .079 | .700 |
| Pairwise Deletion | .015 | .118 | .292 |
| Mean Substitution | .023 | .160 | .449 |
| Bucks Procedure | .016 | .122 | .144 |
| Mean Regression | .018 | .142 | .142 |
| 1 Iteration | .017 | .140 | .073 |
| Iterative Regression | .017 | .140 | .053 |

Note. γ = Kaiser's Gamma. ^adf=4, >300. ^bdf=4, >3000.

*p<.05. **p<.01.

Table 7

Results of Eight-Variable Means Comparison

| Method | Proportion Incomplete | | |
|---|-----------------------|--------|----------|
| | 10% | 20% | 40% |
| <u>Small samples^a, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | .095 | .155 | .909 |
| Pairwise Deletion | .077 | .020 | .423 |
| Mean Substitution | .082 | .023 | .578 |
| Bucks Procedure | .080 | .018 | .479 |
| Mean Regression | .082 | .021 | .606 |
| 1 Iteration | .082 | .022 | .618 |
| Iterative Regression | .082 | .022 | .629 |
| <u>Small samples^a, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .049 | .203 | .285 |
| Pairwise Deletion | .055 | .183 | .215 |
| Mean Substitution | .060 | .218 | .333 |
| Bucks Procedure | .102 | .220 | .214 |
| Mean Regression | .056 | .217 | .319 |
| 1 Iteration | .056 | .221 | .323 |
| Iterative Regression | .056 | .243 | .373 |
| <u>Large samples^b, $\gamma = 0.2$</u> | | | |
| Listwise Deletion | .366 | 2.560* | 10.649** |
| Pairwise Deletion | .162 | .197 | .274 |
| Mean Substitution | .156 | .175 | .346 |
| Bucks Procedure | .121 | .167 | .424 |
| Mean Regression | .167 | .186 | .323 |
| 1 Iteration | .168 | .187 | .344 |
| Iterative Regression | .168 | .187 | .357 |
| <u>Large samples^b, $\gamma = 0.4$</u> | | | |
| Listwise Deletion | .377 | .483 | 3.838** |
| Pairwise Deletion | .645 | .599 | 1.490 |
| Mean Substitution | .706 | .687 | 2.258* |
| Bucks Procedure | .755 | .653 | 1.542* |
| Mean Regression | .747 | .750 | 2.248* |
| 1 Iteration | .753 | .750 | 2.313* |
| Iterative Regression | .779 | .779 | 2.673** |

Note. γ = Kaiser's Gamma. ^adf=8, >300. ^bdf=8, >3000.

* $p < .05$. ** $p < .01$.

covariance matrix, decreased ($p < .05$). The only methods that consistently reproduced the target sample covariance matrix effectively ($p > .05$) when the proportion of incomplete cases was high were listwise and pairwise deletion. Pairwise deletion effectively ($p > .05$) reproduced the target sample covariance matrix when the sample size was large or there were eight variables or the average intercorrelation was high. Pairwise deletion always reproduced the target sample variable means effectively ($p > .05$). Listwise deletion effectively ($p > .05$) reproduced the target sample covariance matrix and variable means whenever there were four variables or the sample size was small.

Table 8 and Table 9 summarize the effectiveness of each method of handling missing values. All conditions are listed as columns. Each method that effectively ($p > .05$) reproduced the target sample covariance matrix is listed in these columns (Table 8). Each method that effectively ($p > .05$) reproduced the target sample variable means is listed in identical columns in Table 9.

Table 8
Summary Table Identifying Methods that Successfully Reproduced the Target Sample Covariance Matrix in the Experimental Conditions

| 4 Variables | | | | 8 Variables | | | | | | | |
|--------------------------|--------|----------------|--------|----------------|--------|----------------|--------|----------------|--------|----------------|--------|
| 10% Incomplete | | 20% Incomplete | | 40% Incomplete | | 10% Incomplete | | 20% Incomplete | | 40% Incomplete | |
| N=200 | N=2000 | N=200 | N=2000 | N=200 | N=2000 | N=200 | N=2000 | N=200 | N=2000 | N=200 | N=2000 |
| Average Intercorrelation | | | | | | | | | | | |
| 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 |
| L | L | L | L | L | L | L | L | L | L | L | L |
| P | P | P | P | P | P | P | P | P | P | P | P |
| MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS |
| BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP |
| MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 |
| IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT |

Note, L = Listwise, P = Pairwise, MS = Mean Substitution, BP = Buck's Procedure, MR = Mean Regression, 11 = 1-Iteration Regression, IT = Iterative Regression.

Table 9
 Summary Table Identifying Methods that Successfully Reproduced the Target Sample Variable Means
 in the Experimental Conditions

| 4 Variables | | | 8 Variables | | | | | | | | | | | | | | | | | |
|--------------------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|----------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|----|----|----|
| 10% Incomplete | 20% Incomplete | 40% Incomplete | 10% Incomplete | 20% Incomplete | 40% Incomplete | 10% Incomplete | 20% Incomplete | 40% Incomplete | | | | | | | | | | | | |
| N=200 | N=200 | N=2000 | N=200 | N=2000 | N=2000 | N=200 | N=200 | N=2000 | | | | | | | | | | | | |
| Average Intercorrelation | | | | | | | | | | | | | | | | | | | | |
| 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | 0.2 | 0.4 | | | |
| L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | L | | |
| P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | P | | |
| MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | MS | | |
| BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | BP | |
| MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | MR | |
| 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | |
| IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT | IT |

Note. L = Listwise, P = Pairwise, MS = Mean Substitution, BP = Buck's Procedure, MR = Mean Regression, 11 = 1-iteration Regression, IT = Iterative Regression.

CHAPTER V

Discussion and Conclusions

As the proportion of incomplete cases increased, methods of handling missing values decreased in effectiveness ($p < .05$). If a method was effective ($p > .05$) with 40% incomplete cases, it was also effective with 20% and 10%. If a method was ineffective ($p < .05$) with 20% incomplete cases, it was ineffective with any higher proportion incomplete. Consequently, 40% incomplete cases provided a more stringent test of the effectiveness of each method to handle missing values (Little & Su, 1987).

Since increased sample size increases the power of a test to detect differences, it should not be surprising to find that more methods failed ($p < .05$) to reproduce either the target sample covariance matrix or the variable means as the sample size increased. However, prior studies have found both listwise and pairwise deletion more effective with large samples (Haitovsky, 1968; Kim & Curry, 1977) and less effective than other methods with small samples (Afifi & Elashoff, 1967; Chan & Dunn, 1972; Gleason & Staelin, 1975). In this study, pairwise deletion was the only method that adequately ($p > .05$) reproduced the target sample covariance matrix and variable means whenever the sample size was large.

The effectiveness of pairwise deletion in this study is in contrast to the results reported by Haitovsky (1968) in which he found listwise deletion superior to pairwise deletion with large sample sizes (1000). Since the sample size was 2000 in this study, one possible reason for a discrepancy may be the difference in sample size. Kim and Curry (1977), however, found pairwise deletion better than listwise deletion with a sample size of 1000. They concluded that Haitovsky's (1968) model was not typical of sociological data and thus could be the source of the discrepancy.

Glasser (1964) reasoned that as sample size increased, approaching infinity, the error produced by the use of pairwise deletion became negligible. This study supports this contention and is in agreement with Kim and Curry's (1977) study. This finding indicates that while increased sample size is more beneficial to pairwise deletion than the other methods, sample size is not as important as some other factors in determining the usefulness of listwise deletion.

Mean substitution has been found more effective than listwise and pairwise deletion with a small sample size (less than 70) and a high proportion of missing values (Afifi & Elashoff, 1967; Chan & Dunn, 1972; Gleason and Staelin, 1975). Since the smallest sample size in this study was 200, and the largest proportion of missing values

is 18% (Table 3), it is not known how effective mean substitution would have been if the sample size was smaller or the proportion of missing values larger.

The regression methods of handling missing data have been found superior to listwise deletion when the average intercorrelation was 0.2 if the number of variables was greater than five (Timm, 1970), or superior to listwise and pairwise deletion if the average inter-correlation was high with as few as two variables (Beale & Little, 1975; Chan & Dunn, 1972; Chan et al., 1976; Timm, 1970). Listwise deletion was superior to one of the regression methods only when there were two variables and low average intercorrelation (Timm, 1970). These studies suggest that the regression methods of handling missing values are more effective than listwise and pairwise deletion; (1) as the proportion of incomplete cases increases, (2) as the average intercorrelation increases, and (3) as the number of variables increases. On the other hand, listwise deletion should be more effective than the regression methods only with two variables and low average intercorrelation. Since the smallest number of variables in this study is four, listwise deletion should never be more effective than the regression methods.

In this study, the only methods that consistently reproduced the target sample covariance matrix or mean

vector effectively ($p > .05$) when the proportion of incomplete cases was high were listwise and pairwise deletion. And, these methods were consistently effective only when qualified by other conditions. Pairwise deletion was effective ($p > .05$) whenever the sample size was large or there were eight variables or the average intercorrelation was high. Listwise deletion was effective ($p > .05$) whenever there were four variables or the sample size was small.

When there were four variables, the covariance matrix produced by the use of listwise deletion was consistently more effective ($p > .05$) in reproducing the target sample covariance matrix than any of the other methods (Table 4). The second most consistently effective method was pairwise deletion which failed to reproduce the target sample covariance matrix only with 40% incomplete cases, small sample size and 0.2 average intercorrelation. The regression methods were consistently ineffective ($p < .05$) when 40% of the cases were incomplete. When sample size was large, the regression methods and mean substitution failed ($p < .05$) to reproduce the target sample covariance matrix with 20% of the cases incomplete when the average intercorrelation was 0.4 and with only 10% incomplete cases when the average intercorrelation was 0.2.

The proportion of missing values accompanying each proportion of incomplete cases was remarkably consistent

when there were four variables. The only difference noted was at 40% incomplete cases. With low correlation, the proportion of missing values was 17% with small samples and 18% with large samples. With high correlation, the proportion of missing values was 16% with small samples and 17% with large samples. The small difference (1%) in proportion of missing values is not sufficient to account for the discrepancy between the results from this study and those of Chan and Dunn (1972), Timm (1970), or Beale and Little (1975). Rather, the distribution of the missing values within the incomplete cases may best explain this inconsistency.

When the average intercorrelation was 0.2 with four variables (Table A-1), 63% of the missing population were missing on variables X3 and X4. These variables had the highest zero order correlation in the matrix (Table 1). When the average intercorrelation was 0.4 with four variables (Table A-3), 60% of the cases in the missing population were missing on variables X3 and X4 jointly. The highest zero order correlation in this correlation matrix was between these two variables (Table 1).

For regression to predict a value effectively, at least one variable must be highly correlated with the dependent variable. In these samples the variable that is most highly correlated with X3 is X4. Since these variables are missing

jointly in more than 59% of the incomplete cases, the estimates produced by Buck's procedure in this instance are produced by regression of the variable with a missing value (X3 or X4) on X1 and X2 only. The estimate thus constructed is based on only two variables rather than three. Neither of the two variables on which the estimate is based is the one most highly correlated with the variable containing the missing value. This situation may have caused the regression methods to be less effective than pairwise deletion.

Haitovsky (1968) suggested that if the pattern of missing values was non-random, assigning a value to the missing entry would work best. This would imply that the use of mean substitution in the initial correlation matrix should produce better results with non-randomly missing data. Prior studies by Chan and Dunn (1972) and Beale and Little (1975) have shown that if the missing values are missing randomly, mean regression and iterative regression are very effective methods of handling missing values. Therefore, whatever the pattern of missing values, mean regression and the iterative regression procedures should be effective. The results of this study, however, were contradictory.

When using the mean regression method, the initial estimate for the missing value was the mean. Regression

equations for estimating missing values were developed using this estimate. When the individual estimate for a missing value in a case was produced, the previously used estimates were used to predict the new estimate. When the variable (X3) most highly correlated with the variable (X4) for which an estimate was being calculated also contained a missing value on this case, the largest contribution to the individual estimate (X4) was made by an estimated value (the mean of X3). As this procedure was iterated this became progressive. In these two samples, since most of the incomplete cases contained jointly missing values from the most highly correlated variables, as the proportion of incomplete cases increased, the regression estimates became progressively worse.

When the average intercorrelation increased, the regression methods were more effective ($p > .05$) than when the average intercorrelation was low. This is in accordance with prior studies. Even though jointly missing values caused the regression methods to fail to reproduce the target sample covariance matrix when the proportion of incomplete cases increased, the two variables that were not missing were more highly correlated with those that were missing and thus were contributing more to their estimation. This is seen in comparing the average intercorrelation with

40% incomplete cases in small samples and with 10% incomplete cases in large samples.

Pairwise deletion failed ($p < .05$) to reproduce the four-variable target covariance matrix in small samples with 40% incomplete cases, and an average intercorrelation of 0.2, while listwise deletion retained the target sample covariance matrix structure. However, there was little difference in their actual performance. With an increase in sample size, both methods performed adequately.

When the number of variables increased to eight (Table 5), more variables were available for the regression methods to use in estimating a missing value. Consequently, these methods should have been more effective. In this study, listwise deletion was not effective ($p < .05$) in reproducing the target sample variable means when there were eight variables and 40% incomplete cases. When the average intercorrelation was 0.2, listwise deletion could not reproduce the target sample covariance matrix. Under these conditions, the regression methods adequately reproduced the covariance matrix. When the average intercorrelation increased, listwise deletion adequately reproduced the covariance matrix, but the regression methods did not.

Kim and Curry (1977) indicated that the effectiveness of listwise deletion deteriorated as the number of variables increased. It appeared that as the number of variables

increased, the proportion of incomplete cases increased. Because listwise deletion is based only on complete cases, fewer cases were available for use. Since listwise deletion is the only method in this study that is based on complete cases rather than missing values, it is more seriously affected by incomplete cases than the other methods.

When the average intercorrelation increased to 0.4, listwise deletion retained the target sample covariance. The regression methods, however, did not perform adequately. Under these conditions, the proportion of missing values for 40% incomplete cases was 15%. When the average intercorrelation was 0.2 with 40% incomplete cases, only 7% of the values were missing. This could account partially for the ineffectiveness of the regression methods. However, there are other reasons also.

When the correlation was 0.4 with eight variables (Table A-4), 52% of the missing population were missing on variables X5, X6, X7, and X8 simultaneously. The zero order correlations among these variables were the highest in this correlation matrix (Table 2). Therefore, missing value estimates in Buck's procedure were produced by regression of the variable with a missing value (X5, X6, X7, or X8) on X1, X2, X3, and X4. The resulting estimate was, therefore, based on four variables instead of seven. Moreover, the variables

used were not highly correlated with the variable being estimated. Since more than 50% of the incomplete cases had these variables missing simultaneously, as the proportion of incomplete cases increased these estimates were progressively worse.

When X5, X6, X7, and X8 were missing simultaneously and the mean regression and iterative regression methods were used, estimates for each missing value were based on one estimated value that contributed most to the variable, two other estimated values, and four actual values (the first four variables). Mean regression and the iterative regression methods were progressively less effective when the average intercorrelation was high as the proportion of incomplete cases increased.

Although listwise deletion reproduced the target sample covariance matrix adequately ($p > .05$) with eight variables and 0.4 average intercorrelation in large samples, it failed to adequately ($p < .05$) reproduce the variable means of the target sample in eight variable situations with large samples at both levels of average intercorrelation. The univariate comparison of the means producing significant differences showed that when the average intercorrelation was low, all the means except those of X1 and X2 were significantly different from the target. When the average intercorrelation was high, all of the means were different.

The distinguishing characteristic between the first two variables and the last two variables in each four variable sample is who initiated contact (parent or school). When the school initiated contact with the parent (variables X1 and X2), there is no evidence of any association in missing values in either sample. When the parent initiated contact with the school (variables X3 and X4), more than 59% of the incomplete cases in the missing population contained missing values on these variables jointly. The reason why some questions when asked by the school (Question 57, parts A-G) were answered but ignored when the parent initiated contact with the school (Question 58, parts A-F) is not known. Possibly the parents completing this survey felt this was redundant or did not comprehend the difference between these two sections. Had there been other questions between questions 57 and 58, possibly both would have been answered. Or, if question 58 had preceded question 57, possibly the parts of question 58 would have been answered and not those of question 57.

The reason why the last four variables (standardized test scores) in the eight variable sample with 0.4 average intercorrelation were missing simultaneously is unknown. It is possible that the school may have refused to report the standardized scores, or the student may have been absent on the dates the standardized tests were administered.

However, in all the samples, except for ones with low average intercorrelation and eight variables, over 50% of the incomplete cases contained jointly missing values on the most highly correlated variables in the sample. Since the incomplete cases were selected from those existing in the NELS-88 data base, this suggests that the occurrence of jointly missing values can be expected when using NELS-88. Since this phenomena was only noted when using variables within the same question, reasearchers using variables that are subsets of a particular question need to be particularly alert to this problem.

When faced with missing data, researchers have two choices: deletion or imputation. Of the two deletion methods, pairwise deletion was more effective when the sample size was large. It has the added advantage of retaining all known data. It simply does not delete observed values. If the jointly missing values were to be used as the dependent variable in a study (such as the standardized test scores), the researcher may prefer to eliminate cases that do not contain these values. In this instance, the choice would be listwise deletion. If the researcher wishes to retain all available information on the cases, pairwise deletion would be the method of choice. Whenever imputation of missing values is desired, an imputation method would be used. However, none of the

imputation methods, in this study, were found effective in retaining the covariance structure when a large proportion of cases were incomplete.

The most surprising result of this study was the relatively strong performance of pairwise deletion. Concerning pairwise deletion, Raymond and Roberts (1987, p15) said "...in instances in which the data were missing in a systematic fashion, it has given misleading results." They further add that it has never been the preferred method. In this study, pairwise deletion failed to reproduce the target sample covariance matrix in only one instance. When the sample size was large, it was always the preferred method.

Conclusion

The regression methods, in this study, have not performed as predicted from prior studies. This discrepancy has been caused by using the proportion of incomplete cases instead of the proportion of missing values as a function of missing data and by the nature of the incomplete cases.

This study was based on one sample in each condition, a limited number of variables, and two levels of average intercorrelation. As such, it is limited in the conclusions that can be drawn or the recommendations that can be made. Keeping these limitations in mind, the following procedures are recommended to the users of NELS-88.

The results of this study suggest that when the proportion of incomplete cases and the sample size are small, all methods used in this study were effective in retaining covariance structure. As the proportion of incomplete cases and sample size increased methods performed differently. Therefore, the selection of a missing data handling method for NELS-88 data must rest on sample size. If the sample size is large (2000), pairwise deletion is more suitable. In small samples, the researcher must determine the proportion of incomplete cases. If the proportion of incomplete cases as well as sample size are small, any of the methods can do the job. If the proportion of incomplete cases is large, the number of variables becomes the crucial factor. With a small number of variables (less than five), listwise deletion has been found effective. When the number of variables is large, pairwise deletion is the choice. The regression methods are not recommended because they are more time-consuming and did not demonstrate any advantage over the other methods.

Further research is needed to determine the effects of the pattern of missing values on the effectiveness of various missing-data-handling methods. Research is also needed to explore the properties of imputation methods when a large proportion of the incomplete cases contain more than one value missing.

References

- Afifi, A.A. & Elashoff, R.M. (1967). Missing observations in multivariate statistics II: Point estimation in simple linear regression. Journal of the American Statistical Association, 62, 10-29.
- Anderson, A.B., Basilevsky, A. & Hum, D. P. J. (1983). Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of Survey Research (pp. 415-494). San Diego: Academic Press Inc.
- Beale, E.M.L. & Little, R.J.A. (1975). Missing values in multivariate analysis. Journal of the Royal Statistical Society, 37, 129-145.
- Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. Journal of the Royal Statistical Society, 22, 302-306.
- Chan, L.S. & Dunn, O.J. (1972). The treatment of missing values in discriminant analysis--1. The sampling experiment. Journal of the American Statistical Association, 67, 473-477.
- Chan, L.S., Gilman, J.A., & Dunn, O.J. (1976). Alternative approaches to missing values in discriminant analysis. Journal of the American Statistical Association, 71, 842-844.

- Cohen, J. & Cohen, P. (1983). Missing data. In J. Cohen & P. Cohen, Applied Multiple Regression/ Correlation Analysis for the Behavioral Sciences (pp. 275-300). Hillsdale, N.J.: Lawrence Erlbaum Associates, Publishers.
- Dempster, A.P., Laird, N.W., & Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, B, 39, 1-38.
- Frane, J.W. (1976). Some simple procedures for handling missing data in multivariate analysis. Psychometrika, 41, 409-415.
- Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. Journal of the American Statistical Association, 59, 834-844.
- Gleason, T.C. & Staelin, R. (1975). A proposal for handling missing data. Psychometrika, 40, 229-251.
- Greenlees, J.S., Reece, W.S., & Ziechang, K.D. (1982). Imputation of missing values when the probability of response depends on the variable being imputed. Journal of the American Statistical Association, 77, 251-261.
- Haitovsky, Y. (1968). Missing data in regression analysis. Journal of the Royal Statistical Society, B, 30, 67-82.

- Hayduk, L.A. (1989). More and better. In L.A. Hayduk, Structural Equation Modeling with LISREL Essentials and Advances, (pp. 276-286). Baltimore: Johns Hopkins University Press.
- Hertel, B.R. (1976). Minimizing error variance introduced by missing data in survey analysis. Sociological Methods & Research, 4, 459-474.
- Ingels, S.J., Abraham, S.Y., Karr, R., Spencer, B.D., Frankel, M.L., & Owings, J.A. (1990a). National Education Longitudinal Study of 1988 Base Year: Student Component Data File User's Manual. U.S. Department of Education, Office of Educational Research and Improvement, NCES 90-464, Washington DC: U.S. Government Printing Office.
- Ingels, S.J., Abraham, S.Y., Rasinski, K.A., Karr, R., Spencer, B.D., Frankel, M.L., & Owings, J.A. (1990b). National Education Longitudinal Study of 1988 Base Year: Parent Component Data File User's Manual. U.S. Department of Education, Office of Educational Research and Improvement, NCES 90-464, Washington DC: U.S. Government Printing Office.
- Jackson, E.C. (1968). Missing values in linear multiple discriminant analysis. Biometrics, 24, 835-844.
- Joreskog, K.G. & Sorbom, D. (1988). Lisrel 7 A guide to the program and applications (2nd ed.). Chicago: SPSS Inc.

- Kaiser, J. (1990, June). The robustness of regression and substitution by mean methods in handling missing values. Paper presented at the 22nd Annual Conference on Statistics, Tours, France.
- Kaiser, J. (1983). The effectiveness of hot-deck procedures in small samples. Proceedings of the Section on Survey Research, American Statistical Association 1983. 523-528.
- Kaiser, J. & Tracy, D.B. (1988). Estimation of missing values by predicted score. Proceedings of the Section on Survey Research, American Statistical Association 1988. 631-635.
- Kalton, G. & Kish, L. (1981). Two efficient random imputation procedures. Proceedings of the Section on Survey Research, American Statistical Association 1981. 146-151.
- Keith, T.Z., Bickley, P., Keith, P.B., Trivett, P.F., Singh, K., Troutman, G.C. (1992, March). Does parental involvement raise eighth grade achievement? Evidence from the National Education Longitudinal Study of 1988. Paper presented at the annual meeting of the National Association of School Psychologists, Knoxville, TN.
- Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods & Research, 6, 215-240.

- LeBlond, G.T., & Cobb, D.F. (1985). Using 1-2-3 [Computer program manual - 2nd edition]. Indianapolis: Que™ Corporation.
- Little, R.J.A., & Rubin, D.R. (1987). Statistical Analysis with Missing Data. New York: John Wiley & Sons.
- Little, R.J.A., & Smith, P.J. (1983). Multivariate edit and imputation for economic data. Proceedings of the Section on Survey Research Methods American Statistical Association 1983. 518-522.
- Little, R.J.A., & Su, H.L. (1987). Missing-data adjustments for partially-scaled variables. Proceedings of the Section on Survey Research Methods American Statistical Association 1987. 644-649.
- Norusis, M.J. (1990). SPSS/PC+™ 4.0 Base Manual [Computer program manual]. Chicago: SPSS Inc.
- Norusis, M.J. (1988a). SPSS/PC+ Advanced Statistics™ V2.0 [Computer program manual]. (pp103-151). Chicago: SPSS Inc.
- Norusis, M.J. (1988b). SPSS-X Introductory Statistics Guide: Release 3 [Computer program manual]. (pp 107-108). Chicago: SPSS Inc.
- Page, E.B. & Keith, T.Z. (1981). Effects of U.S. private schools: A technical analysis of two recent claims. Educational Researcher, 10, 7-17.

- Raymond, M.R. & Roberts, D.M. (1987). A comparison of methods for treating incomplete data in selection research. Educational and Psychological Measurement, 47, 13-26.
- Rummel, R.J. (1970). Handling missing data. In R.J. Rummel, Applied Factor Analysis (pp. 258-265). Evanston: Northwestern University Press.
- Tatsuoka, M.M. (1988). Test of equality of population covariance matrices. In M.M. Tatsuoka, Multivariate Analysis Techniques for Educational and Psychological Research (2nd. ed.), (pp. 98-107). New York: Macmillan Publishing Company.
- Timm, N.H. (1970). The estimation of variance-covariance and correlation matrices from incomplete data. Psychometrika, 35, 417-437.
- Ward, Jr., T.J. & Clark III, H.T. (1991). A reexamination of public-versus private-school achievement: the case for missing data. Journal of Educational Research, 84, 153-163.
- Wilks, S.S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. Annals of Mathematical Statistics, 3, 163-195.

Witta L. & Kaiser, J. (1991, November). Four methods of handling missing data with GSS-84. Paper presented at the meeting of the Mid-South Educational Research Association, Lexington, KY

Wofle, L.M. (1985). Postsecondary educational attainment among whites and blacks. American Educational Research Journal, 22, 501-525.

Appendix A

Missing Data Characteristics

Tables 1-4 in this appendix contain the frequency missing for each variable and each variable combination. The first column lists the variable or variable combination that is missing. The column labeled frequency shows the number of times that variable singly or variable combination is missing. The first percent column shows the proportion of incomplete cases for which this variable or variable combination accounts. The second frequency column shows the proportion of incomplete cases accounted for by that combination after removal of those cases having more than half of the variable values missing.

Tables included in this appendix are:

- Table 1: Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Four Variables
- Table 2: Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables
- Table 3: Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Four Variables
- Table 4: Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Eight Variables

Table 1

Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Four Variables

| Incomplete Cases | | | |
|------------------|-----------|----------------------|----------------------|
| Missing Value | Frequency | Percent ^a | Percent ^b |
| >2 Missing | 162 | 9.4 | .0 |
| X1 | 66 | 3.8 | 6.4 |
| X2 | 52 | 3.0 | 5.1 |
| X3 | 75 | 4.4 | 7.3 |
| X4 | 63 | 3.7 | 6.1 |
| X1 X2 | 81 | 4.7 | 7.9 |
| X1 X3 | 5 | .3 | .5 |
| X1 X4 | 18 | 1.0 | 1.8 |
| X2 X3 | 5 | .3 | .5 |
| X2 X4 | 10 | .6 | 1.0 |
| X3 X4 | 650 | 37.8 | 63.4 |
| Nonresponse | 531 | 30.9 | .0 |
| | ----- | ----- | ----- |
| TOTAL | 1718 | 100.0 | 100.0 |

Note. Listed by missing value combination before and after removal of cases with more than two missing values.

^aPrior to removal of cases containing more than two missing values. ^bAfter removal of cases containing more than two missing values.

Table 2 (Part 1)

Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables

| Missing Value | Incomplete Cases | | |
|---------------|------------------|----------------------|----------------------|
| | Frequency | Percent ^a | Percent ^b |
| >5 Missing | 198 | 4.1 | .0 |
| X1 | 355 | 7.3 | 7.7 |
| X2 | 116 | 2.4 | 2.5 |
| X3 | 177 | 3.6 | 3.8 |
| X4 | 1154 | 23.7 | 25.0 |
| X5 | 354 | 7.3 | 7.7 |
| X6 | 74 | 1.5 | 1.6 |
| X7 | 299 | 6.1 | 6.5 |
| X8 | 717 | 14.7 | 15.5 |
| X1 X2 | 13 | .3 | .3 |
| X1 X3 | 14 | .3 | .3 |
| X1 X4 | 29 | .6 | .6 |
| X1 X5 | 14 | .3 | .3 |
| X1 X6 | 3 | .1 | .1 |
| X1 X7 | 17 | .3 | .4 |
| X1 X8 | 11 | .2 | .2 |
| X2 X3 | 4 | .1 | .1 |
| X2 X4 | 3 | .1 | .1 |
| X2 X5 | 10 | .2 | .2 |
| X2 X6 | 3 | .1 | .1 |
| X2 X7 | 15 | .3 | .3 |
| X2 X8 | 4 | .1 | .1 |
| X3 X4 | 19 | .4 | .4 |
| X3 X5 | 15 | .3 | .3 |
| X3 X6 | 1 | .0 | .0 |
| X3 X7 | 11 | .2 | .2 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

Table 2 (Part 2)

Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables

| Missing Value | Incomplete Cases | | |
|---------------|------------------|----------------------|----------------------|
| | Frequency | Percent ^a | Percent ^b |
| X3 X8 | 10 | .2 | .2 |
| X4 X5 | 70 | 1.4 | 1.5 |
| X4 X6 | 5 | .1 | .1 |
| X4 X7 | 32 | .7 | .7 |
| X4 X8 | 35 | .7 | .8 |
| X5 X6 | 1 | .0 | .0 |
| X5 X7 | 14 | .3 | .3 |
| X5 X8 | 16 | .3 | .3 |
| X6 X7 | 454 | 9.3 | 9.8 |
| X6 X8 | 4 | .1 | .1 |
| X7 X8 | 10 | .2 | .2 |
| X1 X2 X3 | 1 | .0 | .0 |
| X1 X2 X4 | 2 | .0 | .0 |
| X1 X2 X5 | 4 | .1 | .1 |
| X1 X2 X6 | 3 | .1 | .1 |
| X1 X2 X7 | 11 | .2 | .2 |
| X1 X3 X4 | 1 | .0 | .0 |
| X1 X3 X5 | 4 | .1 | .1 |
| X1 X3 X7 | 3 | .1 | .1 |
| X1 X3 X8 | 3 | .1 | .1 |
| X1 X4 X5 | 6 | .1 | .1 |
| X1 X4 X7 | 3 | .1 | .1 |
| X1 X4 X7 | 3 | .1 | .1 |
| X1 X5 X6 | 1 | .0 | .0 |
| X1 X5 X7 | 3 | .1 | .1 |
| X1 X5 X8 | 5 | .1 | .1 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

Table 2 (Part 3)

Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables

| Missing Value(s) | Incomplete Cases | | |
|------------------|------------------|----------------------|----------------------|
| | Frequency | Percent ^a | Percent ^b |
| X1 X6 X7 | 11 | .2 | .2 |
| X2 X3 X4 | 1 | .0 | .0 |
| X2 X3 X6 | 3 | .1 | .1 |
| X2 X3 X7 | 1 | .0 | .0 |
| X2 X3 X8 | 1 | .0 | .0 |
| X2 X4 X5 | 1 | .0 | .0 |
| X2 X4 X7 | 2 | .0 | .0 |
| X2 X4 X8 | 1 | .0 | .0 |
| X2 X5 X7 | 2 | .0 | .0 |
| X2 X5 X8 | 2 | .0 | .0 |
| X2 X6 X7 | 156 | 3.2 | 3.4 |
| X2 X7 X8 | 1 | .0 | .0 |
| X3 X4 X5 | 11 | .2 | .2 |
| X3 X4 X7 | 1 | .0 | .0 |
| X3 X5 X7 | 3 | .1 | .1 |
| X3 X5 X8 | 5 | .1 | .1 |
| X3 X6 X7 | 8 | .2 | .2 |
| X4 X5 X7 | 4 | .1 | .1 |
| X4 X5 X8 | 5 | .1 | .1 |
| X4 X6 X7 | 38 | .8 | .8 |
| X4 X6 X8 | 1 | .0 | .0 |
| X4 X7 X8 | 1 | .0 | .0 |
| X5 X6 X7 | 14 | .3 | .3 |
| X5 X7 X8 | 1 | .0 | .0 |
| X6 X7 X8 | 23 | .5 | .5 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

Table 2 (Part 4)

Frequency of Population Incomplete Cases for Low Average Intercorrelation (0.2) and Eight Variables

| Missing Value(s) | Incomplete Cases | | |
|------------------|------------------|----------------------|----------------------|
| | Frequency | Percent ^a | Percent ^b |
| X1 X2 X4 X7 | 1 | .0 | .0 |
| X1 X2 X5 X7 | 3 | .1 | .1 |
| X1 X2 X6 X7 | 106 | 2.2 | 2.3 |
| X1 X3 X4 X5 | 5 | .1 | .1 |
| X1 X3 X6 X7 | 1 | .0 | .0 |
| X1 X4 X5 X6 | 1 | .0 | .0 |
| X1 X4 X5 X8 | 1 | .0 | .0 |
| X1 X4 X6 X7 | 1 | .0 | .0 |
| X1 X4 X6 X8 | 1 | .0 | .0 |
| X1 X5 X6 X7 | 4 | .1 | .1 |
| X2 X3 X4 X5 | 1 | .0 | .0 |
| X2 X3 X5 X6 | 1 | .0 | .0 |
| X2 X3 X6 X7 | 5 | .1 | .1 |
| X2 X3 X6 X8 | 1 | .0 | .0 |
| X2 X4 X6 X7 | 18 | .4 | .4 |
| X2 X5 X6 X7 | 10 | .2 | .2 |
| X2 X6 X7 X8 | 7 | .1 | .2 |
| X3 X4 X5 X7 | 2 | .0 | .0 |
| X3 X4 X6 X7 | 2 | .0 | .0 |
| X4 X5 X6 X7 | 5 | .1 | .1 |
| X4 X5 X7 X8 | 1 | .0 | .0 |
| X4 X6 X7 X8 | 2 | .0 | .0 |
| X5 X6 X7 X8 | 1 | .0 | .0 |
| NONRESPONSE | 55 | 1.1 | .0 |
| | ----- | ----- | ----- |
| TOTAL | 4866 | 100.0 | 100.0 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

Table 3

Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Four Variables

| Incomplete Cases | | | |
|------------------|-----------|----------------------|----------------------|
| Missing Value(s) | Frequency | Percent ^a | Percent ^b |
| >2 Missing | 96 | 5.4 | .0 |
| X1 | 45 | 2.5 | 4.0 |
| X2 | 95 | 5.4 | 8.4 |
| X3 | 53 | 3.0 | 4.7 |
| X4 | 85 | 4.8 | 7.5 |
| X1 X2 | 96 | 5.4 | 8.5 |
| X1 X3 | 6 | .3 | .5 |
| X1 X4 | 7 | .4 | .6 |
| X2 X3 | 3 | .2 | .3 |
| X2 X4 | 58 | 3.3 | 5.2 |
| X3 X4 | 678 | 38.4 | 60.2 |
| Nonresponse | 545 | 30.8 | .0 |
| | ----- | ----- | ----- |
| TOTAL | 1767 | 100.0 | 100.0 |

Note. Listed by missing value combination before and after removal of cases with more than two missing values.

^aPrior to removal of cases containing more than two missing values. ^bAfter removal of cases containing more than two missing values.

Table 4 (Part 1)

Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Eight Variables

| Missing Value(s) | Incomplete Cases | | |
|------------------|------------------|----------------------|----------------------|
| | Frequency | Percent ^a | Percent ^b |
| >5 Missing | 76 | 4.5 | .0 |
| X1 | 36 | 2.1 | 2.2 |
| X2 | 8 | .5 | .5 |
| X3 | 51 | 3.0 | 3.2 |
| X4 | 87 | 5.1 | 5.4 |
| X5 | 30 | 1.8 | 1.9 |
| X6 | 42 | 2.5 | 2.6 |
| X7 | 6 | .4 | .4 |
| X8 | 95 | 5.6 | 5.9 |
| X1 X2 | 10 | .6 | .6 |
| X1 X3 | 9 | .5 | .6 |
| X1 X4 | 24 | 1.4 | 1.5 |
| X2 X3 | 12 | .7 | .7 |
| X2 X4 | 19 | 1.1 | 1.2 |
| X3 X4 | 23 | 1.4 | 1.4 |
| X3 X5 | 2 | .1 | .1 |
| X3 X6 | 1 | .1 | .1 |
| X4 X6 | 1 | .1 | .1 |
| X4 X8 | 2 | .1 | .1 |
| X5 X6 | 2 | .1 | .1 |
| X5 X7 | 9 | .5 | .6 |
| X5 X8 | 6 | .4 | .4 |
| X6 X7 | 3 | .2 | .2 |
| X6 X8 | 4 | .2 | .2 |
| X7 X8 | 46 | 2.7 | 2.8 |
| X1 X2 X3 | 11 | .7 | .7 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

Table 4 (Part 2)

Frequency of Population Incomplete Cases for High Average Intercorrelation (0.4) and Eight Variables

| Incomplete Cases | | | |
|------------------|-----------|----------------------|----------------------|
| Missing Value(s) | Frequency | Percent ^a | Percent ^b |
| X1 X2 X4 | 31 | 1.8 | 1.9 |
| X1 X2 X8 | 1 | .1 | .1 |
| X1 X3 X4 | 16 | .9 | 1.0 |
| X1 X3 X5 | 1 | .1 | .1 |
| X1 X6 X7 | 1 | .1 | .1 |
| X2 X3 X4 | 24 | 1.4 | 1.5 |
| X2 X4 X5 | 1 | .1 | .1 |
| X3 X4 X5 | 1 | .1 | .1 |
| X3 X7 X8 | 1 | .1 | .1 |
| X4 X7 X8 | 1 | .1 | .1 |
| X5 X6 X7 | 1 | .1 | .1 |
| X5 X7 X8 | 7 | .4 | .4 |
| X1 X2 X3 X4 | 148 | 8.7 | 9.2 |
| X1 X2 X4 X6 | 1 | .1 | .1 |
| X1 X2 X4 X8 | 1 | .1 | .1 |
| X1 X4 X7 X8 | 1 | .1 | .1 |
| X2 X3 X5 X8 | 1 | .1 | .1 |
| X2 X5 X7 X8 | 1 | .1 | .1 |
| X5 X6 X7 X8 | 838 | 49.5 | 51.9 |
| | ----- | ----- | ----- |
| TOTAL | 1692 | 100.0 | 100.0 |

Note. Listed by missing value combination before and after removal of cases with more than four missing values.

^aPrior to removal of cases containing more than four missing values. ^bAfter removal of cases containing more than four missing values.

V i t a

Eleanor Lea Witta
Rt 3, Box 124
Abingdon, Virginia 24210

Phone 703-628-1419
Birthdate 12-2-42

Education

Virginia Polytechnic Inst & State U, Blacksburg, Virginia.
Degree: Phd EDRE Graduation: 1992
ETSU, Johnson City, TN - Classes transferred to VA. Tech.
UVA, Charlottesville, Va. - Classes transferred to VA. Tech.
Texas A & M University, College Station, Texas
Degree: Masters EDCI Graduation: 1973
West Virginia University, Morgantown, West Virginia
Degree: BA Psychology Graduation: 1964

Presentations & Publications

Keith, Timothy Z. and Eleanor L. Witta, Hierarchical and Cross-Age Confirmatory Factor Analysis of the WISC-III: What Does It Measure?, Submitted to the Journal of School Psychology, July 1992.
Witta, Lea, "Measuring intelligence with the WISC III: What does it really measure?" Poster presentation for the annual Virginia Tech Graduate Student Research Symposium, Blacksburg, VA, April, 1992. (Mentor - Tim Z. Keith)
Witta, Lea and Javaid Kaiser, "Four methods of handling missing data with the 1984 General Social Survey." Presented to the Mid-South Educational Research Association Annual Conference, Lexington, KY, Nov., 1991.

Work Experience

ETSU, Johnson City, Tennessee
Position: Adjunct Faculty Sept 92 to present
Virginia Tech, Blacksburg, Virginia
Position: Computer Lab Assistant June 91 to Aug 91
Virginia Tech, Blacksburg, Virginia
Position: Internship Survey Anlysis Aug 90 to Dec 90
ETSU, Johnson City, Tennessee
Position: Graduate Research Asst Sept 89 to Dec 89
Tazewell County Schools, Tazewell, Virginia
Position: Earth Science Teacher Aug 78 to June 80
Buchanan County Schools, Grundy, Virginia
Position: Science Teacher Nov 68 to June 74

