Advances in Missing Data Methods and Implications for Educational Research

Chao-Ying Joanne Peng, Indiana University-Bloomington

Michael Harwell, University of Minnesota

Show-Mann Liou

Lee H. Ehman

Indiana University-Bloomington

KEY WORDS: missing data, listwise deletion, pairwise deletion, mean substitution, single imputation, multiple imputation, EM algorithm, educational research

Author Note

Send all correspondence to Chao-Ying Joanne Peng, Department of Counseling and Educational Psychology, School of Education, Room 4050, 201 N. Rose Ave., Indiana University, Bloomington, IN 47405-1006, tel: (812) 856-8337, fax: (812) 856-8333, email: peng@indiana.edu. We wish to thank Gary M. Ingersoll for his permission to use the Adolescent data set.

1. In c:\u\articles\liou\rer\miss.rer.fin.5-6-03 (5/6/03), miss.rer.peng-harwell.1st.doc (3/31/03) or miss.rer.1st.doc (3/31/03), miss.rer.doc (11/15/2002). Short version presented at 2002 CAERDA in miss.caerda.short.doc (3/28/02). Taipei conference (7/10/02) in miss.Taipei.long2.doc or miss.Taipei.short.doc.
3. Previous versions in [miss.trse.peng.doc] or [liou.minor.doc], [liou.missing.doc], or [miss.trse.fin.doc]
Abstract

The impact of missing data on quantitative educational research has been a great concern to methodologists but less so for many educational researchers. In studies where missing data are present, it is common for educational researchers to use ad hoc methods such as listwise or pairwise deletion to deal with the missing data problem. Yet the recent American Psychological Association Task Force on Statistical Inference warned against the use of these methods for dealing with missing data and encouraged researchers to instead employ alternative methods that take into account the conditions under which missing data occurred. These newer methods have been implemented in statistical software (e.g., SPSS® and SAS®); thus, they are accessible to researchers. In this paper, we promote the use of principled statistical methods for treating missing data that employ single- or multiple-imputation of missing values. These methods are illustrated with real-world data sets, and the results are shown to be different from those obtained from listwise deletion method. Statistical assumptions underlying these principled methods are discussed and emphasized. Quality of educational research will be enhanced if (a) researchers explicitly acknowledge missing data problems and the conditions under which they occurred, (b) principled, rather than ad hoc methods, are used to handle missing data, and (c) journal referees and editors incorporate the reporting of appropriate missing data methods in standards used in judging manuscripts for publication.
**Advances in Missing Data Methods and Implications for Education Research**

Consider a scenario in which a freshly minted Ph.D. showed off his bound dissertation to a fellow graduate student who was known among her peers to be a “number cruncher.” As she leafed through the pages of the methods section, she quickly noticed that all missing data were coded zero, instead of being declared as missing. She also saw that all analyses in this dissertation, including factor analysis, multiple regression, discriminant function analyses, etc., were based on all data including the zeros. No one on the dissertation committee caught this mistake. In another scenario, a Ph.D. candidate explaining the results of the data analyses to her dissertation committee noted that the choice of listwise deletion led to the removal of approximately one third of the data. Both the candidate and committee members failed to recognize the possibly devastating effects on inferences due to these missing data. Variations of these scenarios may occur more frequently in educational research than is commonly believed or acknowledged.

Missing data in research studies are the rule rather than the exception. For example, various literatures indicate that a 20% attrition rate has been held as a benchmark for studies related to youth, school-based programs, and clinical research (Hall, 1993; Kellam, Rebok, Ialongo, & Mayer, 1994; Mason, 1999). Many reasons contribute to data missing from research projects: subject’s attrition in pretest-posttest or longitudinal designs, equipment failure or unclear instruction or intrusive questions on a survey. Whatever the reason(s) for missing data, their impact on quantitative research has been a great concern to methodologists.

In survey research, missing data problems are particularly acute, with substantial amounts of missing data frequently appearing (Little & Rubin, 1987). To combat these problems, investigators spend enormous time, funding, and energy to minimize incomplete data or non-
response among respondents (Mason, 1999). Procedures have been devised by statisticians to mitigate inferential or descriptive problems caused by missing data (Dempster, Laird & Rubin, 1977; Glasser, 1964; Rubin, 1987; Schafer, 1997). The development of these principled procedures is a unique feature of the survey research literature. For the past two decades, the statistical properties of these procedures have been studied and made known. And many have been incorporated into widely accessible software since 1990 (Cool, 2000; Kim & Curry, 1977; Mason, 1999; Raymond & Roberts, 1987; Witta, 2000). Even so, an examination of articles published in 11 education journals for the past 5 years reveals that, in studies where missing data were reported and care was taken to treat missing data, an overwhelming majority of these studies used ad hoc methods, such as listwise and pairwise deletion, to handle missing data. Ad hoc methods are characterized by their failure to take into account the mechanism that led to missing data. Yet the recent APA Task Force on Statistical Inference explicitly warned against the use of these methods:

Special issues arise in modeling when we have missing data. The two popular methods for dealing with missing data that are found in basic statistics packages—listwise and pairwise deletion of missing values—are among the worst methods available for practical applications. (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 598)

Newer and principled methods take into consideration conditions under which missing data occurred. In this paper, we promote the use of three such methods, namely, two single-imputation methods and one based on multiple imputation. These methods are illustrated with real-world data sets, and the results are shown to be different from those obtained from listwise deletion. Key issues in guiding the use of these methods are also discussed. Since these methods have been implemented into statistical software (e.g., SPSS® and SAS®) and should be widely
available, their utility by educational researchers is expected to substantially increase. Throughout our illustration and discussion, the focus is on handling cases in which partial information is missing rather than a complete lack of information.


1. Problems Caused by Missing Data and Missing Data Treatments

Why are missing data a problem? The most serious concern is that missing data can introduce bias into estimates derived from a statistical model (Becker & Walstad, 1990; Becker & Powers, 2001; Holt, 1997; Rubin, 1987). For example, it is possible that non-respondents might have different response profiles compared to those who responded completely. Thus, the remaining sample is no longer random or representative of the population from which it was randomly drawn. If the researcher chose to draw their conclusions based solely on those who responded, the conclusions would be biased.

Furthermore, missing data result in a loss of information and statistical power (Anderson, Basilevsky & Hum, 1983; Kim & Curry, 1977). The elimination of subjects with missing information on one or more variables from the statistical analysis in listwise deletion decreases the error degrees of freedom (df) in statistical tests such as the $t$. This decrease in turn leads to reduced statistical power and larger standard errors compared to those obtained from complete random samples (Cohen & Cohen, 1983; Cool, 2000).
Similar loss of df and statistical power occurs with pairwise deletion, where standard deviations, correlations and covariances are calculated on the basis of available data on each variable (Glasser, 1964; Raymond & Robert, 1987). As a result, the sample composition differs from variable to variable, and the population to which the results are generalized is no longer clearly defined.

Another problem with missing data is that they make common statistical methods inappropriate or difficult to apply (Rubin, 1987). For example, when missing data are present in a factorial analysis of variance the design is unbalanced. Consequently, the standard statistical analysis that is appropriate for balanced designs is no longer appropriate under this condition. Even if data are assumed to be missing in a completely random fashion, the proper analysis is complicated. Multivariate statistical methods, as they are programmed into commercial statistical software, are applicable to complete data sets by default.

Finally, valuable resources are wasted as a result of missing data. Time and funding spent on subjects who subsequently leave a study and/or produce missing data represents a loss (Buu, 1999; Holt, 1997). Such loss is a particular concern in longitudinal research, large-scale assessments, high-stake studies, and/or surveys that ask sensitive information or target respondents who are not accustomed to responding to opinions surveys (such as the first generation Hmong immigrants). Efforts to achieve higher response rates and complete profiles from respondents require researchers to allocate additional time and resources to trace cases who failed to respond or those whose responses were incomplete or unusable. These efforts may not always pay off.

2. Missing Data Methods Reported in 11 Education Journals
In order to understand how educational researchers currently deal with missing data, we reviewed the quantitative studies published in 11 education journals from 1998 to the summer of 2002. The 11 journals were *American Educational Research Journal (AERJ)*, *Educational Researcher (ER)*, *Journal of Counseling Psychology (JCP)*, *Journal of Educational Psychology (JEP)*, *Journal of Research in Science Teaching (JRST)*, *Journal of Special Education (JSE)*, *Journal of School Psychology (JSP)*, *The Modern Language Journal (MLJ)*, *Research in Higher Education (RHE)*, *Journal for Research in Mathematics Education (RME)*, and *Theory and Research in Social Education (TRSE)*. These journals were selected because of their emphasis on research, broad coverage of research topics, relevance to sub-fields in education, and reputable editorial policies. We assumed that the research reported in these 11 journals reflected the mainstream topics and research methods used in educational research.

Within the review period, we identified 1087 studies in 918 articles that met our criteria for quantitative research. For each study, the methods, findings, and discussion sections (or their equivalents) and all summary tables reported were examined by two of the authors. Special attention was given to the total sample size, the \( df \) of the test statistics reported, and information regarding authors’ treatment of missing data. The unit of analysis was studies, not articles, as many articles included two or more empirical studies.

Of the 1087 studies, 305 (28%) did not report any missing data problem, 587 (54%) exhibited evidence of missing data, and the remaining 195 (18%) did not provide sufficient information (such as the total sample size, \( df \) of the statistics, means and percentages, etc.) to help us determine if missing data were present. Among the 587 studies that showed evidence of missing data (such as the mismatch between the sample size and error degrees of freedom), 569 (97%) explicitly or implicitly reported dealing with such a problem. Of these 569 studies, 509
(89.5%) used the listwise deletion (LD) method and 43 (7.6%) the pairwise deletion (PD) method. The heavy reliance on listwise or pairwise deletion is probably attributable to the fact that several popular and accessible statistical software programs, such as SPSS® or SAS®, default to listwise or pairwise deletion for handling missing data.

A small number of studies (17 or 2.9%) in the review used other methods for handling missing data (Table 1). A breakdown of the statistics by the year 1999, when the APA Task Report was published, revealed that only four studies (all published in or after 2000) used a newer, more principled method (i.e., EM). These results suggest that educational researchers have not as yet actively applied principled missing data methods in empirical studies. And refereed journals have not yet encouraged authors to steer away from LD or PD methods.

3. Missing Data Mechanisms

Missing data occur to varying degrees and in various patterns (Cohen & Cohen, 1983, pp.275-299). The impact of missing data on the validity of research findings depends on the mechanisms that led to missing data, the pattern of missing data, and the proportion of data missing (Tabachnick & Fidell, 2001, p.58). Each is discussed below.

It has been shown that the mechanism and the pattern of missing data have greater impact on research results than does the amount of data missing (Tabachnick & Fidell, 2001, p.58). Both are critical issues a researcher must address before choosing an appropriate procedure to deal with missing data. According to Little and Rubin (1987), mechanisms that lead to missing data can be classified as: missing completely at random, missing at random, and non-ignorable missing. As defined by Little and Rubin, if the probability of a response depends on neither the observed nor the missing value that could have been collected or recorded, the missing data are missing completely at random.
As an example, consider a hypothetical data comprising measurements of ten subjects on three variables \((Y, X_1, X_2)\) in Table 2A. For purpose of discussion, let’s assume that \(Y\) represented post-test scores, \(X_1\) pre-test scores, and \(X_2\) IQ scores. If the likelihood that \(Y\) is missing is unrelated to the missing value itself, nor with \(X_1\) or \(X_2\), either collected or missing, then \(Y\) is said to be *missing completely at random* (abbreviated as MCAR). Under the MCAR condition, missing data can be treated as a random sub-sample of the potentially complete data, and the missing data mechanism capturing the reasons for missing data can be ignored for sampling-based and likelihood-based inferences (Little & Rubin, 1987, p. 15).

Let \(Y\) denote a data vector composed of two parts: those completely observed and those potentially missing. In other words, \(Y = (Y_{\text{observed}}, Y_{\text{missing}})\). If the probability of \(Y\) being missing does not depend on the missing value itself, but can depend on observed values of \(Y\) or other completely observed variables (\(X\’s\)), then missing data are said to be *missing at random*. Formally defined, the *missing at random* (abbreviated as MAR) assumption states that

\[
\text{Probability } (Y_{\text{missing}} \mid Y, X_j, j=1 \text{ to } k) = \text{Probability } (Y_{\text{missing}} \mid Y_{\text{observed}}, X_j, j=1 \text{ to } k). \tag{1}
\]

Eq. (1) implies that the conditional probability of \(Y\) being missing, given both \(Y\) and \(X_j\) is the same as the conditional probability of missing values on \(Y\), given observed values of \(Y\) and completely observed variables \(X_j\). Using the hypothetical data in Table 2A, MAR means that any student’s missing score on \(Y\) (the post test) could be related to \(X_1\) (pre-test) or \(X_2\) (IQ) but not to the missing \(Y\) score that could have been collected.
MAR is less restrictive than MCAR; thus, MCAR is said to be a special case of MAR. Under the condition of MAR, the missing mechanism is *ignorable* for likelihood-based inferences (Little & Rubin, 1987, p.15). According to Allison (2001, p. 5):

The missing data mechanism is said to be ignorable if (a) the data are MAR and (b) the parameters that govern the missing data process are unrelated to the parameters to be estimated. Ignorability basically means that there is no need to model the missing data mechanism as part of the estimation process. However, special techniques certainly are needed to utilize the data in an efficient manner. Because it is hard to imagine real-world applications where condition (b) is not satisfied, I [i.e., Allison] treat MAR and ignorability as equivalent conditions in this book. Even in the rare situation where condition (b) is not satisfied, methods that assume ignorability work just fine, but you could do even better by modeling the missing data mechanism.

Based on this logic, MAR and ignorability will be treated as interchangeable in this paper.

In contrast to *missing at random*, the missing data are *non-ignorable* if the probability of missing data depends on the missing values themselves. Again, using the hypothetical data in Table 2A, suppose that students missed the post test \( Y \) because they were poorly prepared for the test, hence their scores were likely to be low. In this case, the missing data on \( Y \) would be said to be *non-ignorable*. Unlike the ignorable case, the missing data mechanism must be specified by the researcher and incorporated into the data analysis in order to produce unbiased parameter estimates, a formidable task. Understandably, *non-ignorable* missing data in educational research are most likely to occur in studies that seek to gather sensitive or personal information. No statistical test exists at the present to examine if this condition is met. Pilot testing of the instrument or common sense can sometimes detect this type of missing mechanism. All missing
data methods presented in this paper are applicable under either the MCAR or the MAR condition.

The condition of MCAR may be examined using Little’s multivariate test, which tests whether the MCAR condition is tenable for the data (Little, Roderick, & Schenker, 1995). Whether the MAR condition holds can be examined by a simple $t$-test of mean differences between the group with complete data and that with missing data (Diggle, Liang, & Zeger, 1994; Kim & Curry, 1977; Tabachnick & Fidell, 2001). Both approaches are illustrated with a data set at http://www.spss.com/SPSSBI/SPSS/mva/. However, we caution readers that the results of these tests cannot be interpreted as providing incontrovertible evidence of either MCAR or MAR.

If the pattern of missing data is monotone, then the estimation of parameters in a multivariate distribution can be simplified. A monotone missing data pattern is illustrated in Table 2B in which missing data are progressively more prevalent between the pretest and the post test. For such a monotone missing data pattern, the less restrictive MAR assumption allows for a simplification of the estimation of parameters in a joint, multivariate distribution. Let’s illustrate this point with three variables in Table 2B. For the joint, multivariate distribution of $Y$, $X_1$, and $X_2$, the probability function $f(Y, X_1, X_2)$ can be expressed as a product of the marginal distribution of $X_1$ and $X_2$ multiplied with the conditional distribution of $Y$ given $X_1$ and $X_2$, as in Equation (2):

$$f(Y, X_1, X_2) = f(X_1, X_2) f(Y | X_1, X_2).$$ (2)

By the same token, Equation (3) is also true:
\[ f(X_1, X_2) = f(X_2) f(X_1 | X_2). \] (3)

Parameters in \( f(X_2) \) are estimated from 10 subjects who had complete data on IQ \( (X_2) \).
Parameters in the conditional distribution of the pretest \( (X_1) \) given IQ are estimated from 9 subjects for whom complete data are available. Together, these results can be combined to estimate parameters in the joint distribution of \( X_1 \) and \( X_2 \) in Eq. (3). By the same logic, parameters in the conditional distribution of posttest \( (Y) \) given the pretest \( (X_1) \) and IQ \( (X_2) \) are estimated from 7 subjects with complete information on all three variables. Results from this inference can be combined with results of Eq. (3) to make inferences about parameters in the joint distribution of all variables, as in Equation (2). If MAR assumption holds for missing data, the conditional distribution in both equations is often estimated in the form of regression equations, such as regressing the pretest on IQ in Eq. 3 or regressing the posttest on both pretest and IQ, as in Eq. (2). Thus, the problem of missing data is solved by replacing them with imputed scores, derived from regression equations. More on this approach is presented in the next section, “4. Overview of Missing Data Methods”.

Regarding the question of how large a proportion of missing data can be tolerated by missing data methods, there are no firm guidelines agreed upon by statisticians at present. If only a few data values are missing in a random pattern from a large data set (i.e., the MCAR condition holds), the missing data problem is less serious and almost any procedure for handling missing data yields similar results. However, if a substantial amount of data is missing from a small to moderate sized data set, the problem can be very serious (Cohen & Cohen, 1983; Cool, 2000; Tabachnick & Fidell, 2001).

4. Overview of Five Ad Hoc Missing Data Methods
The history of the development of missing data methods can be divided into three periods (Schafer, 1997). In the first period, prior to 1980, most widely applied methods dealing with missing data were ad hoc. These include LD, PD, mean substitution, simple hot-deck method, and various regression-based methods. They are easy to use, yet typically produce biased results. In the second period, roughly beginning with the publication of Little and Rubin (1987), principled methods, such as the full information maximum likelihood (FIML) and the Expectation-Maximization (EM) algorithm, began to appear. These methods are generally superior to ad hoc methods in that they are statistically efficient and produce parameter estimates with acceptable standard errors. Even though these methods are model-specific and can be difficult to implement, they are viewed as breakthroughs in the history of missing data methods.

The third period in the development of missing data methods began in the late 80’s and early 90’s; it was characterized by the introduction of multiple imputation methods to overcome limitations of single imputation methods. Simulation studies have shown this method to be flexible and to yield smaller standard errors than those obtained by other procedures. Even though the multiple imputation method represents the latest effort by methodologists to deal with missing data, it has not been widely adopted by educational researchers.

In this section, we review five ad hoc methods for handling missing data because of their prevalence in statistical software. These include two described earlier (i.e., listwise and pairwise deletion), along with mean substitution, simple hot-deck, and regression. The strengths and weaknesses of each method are discussed in terms of parameter estimation and hypothesis testing. A summary of features, strengths, and weaknesses of all methods is given in Table 3.

(1) Listwise Deletion (LD)
As noted earlier, LD removes subjects that have missing information on one or more variables from the statistical analysis. As Kim and Curry (1977) show, 59% of the data can be lost using LD if only 10% of the data were eliminated randomly from each variable in a data set with five variables.

LD is the easiest and most common method for handling missing data. It was used in 89.5% of the studies published in 11 education journals between 1998 and the summer of 2002. This popularity is in large part due to the use of LD as the default setting for multivariate and several univariate statistical procedures in popular statistical packages such as SPSS®, SYSTAT®, or SAS®.

Although LD is generally not recommended, it can be safely used if the correlations among four or fewer variables are low to average and only a small proportion of missing values are present (Buck, 1960; Haitovsky, 1968; Timm, 1970). As long as the MCAR assumption holds for missing data, Allison (2001, p.84) asserted that among conventional methods for handling missing data, listwise deletion is the least problematic. Although listwise deletion may discard a substantial fraction of the data, there is no reason to expect bias unless the data are not missing completely at random. In addition, the standard errors also should be decent estimates of the true standard errors. Furthermore, if you are estimating a linear regression model, listwise deletion is quite robust to situations where there are missing data on an independent variable and the probability of missingness depends on the value of that variable. If you are estimating a logistic regression model, listwise deletion can tolerate either nonrandom missingness on the dependent variable or nonrandom missingness on the independent variables (but not both).
In other words, under MCAR, LD does not produce biased estimates and standard statistical procedures are applicable to data. Yet, the loss in statistical power and in the precision of estimation should not be overlooked, nor can they be compensated. If the MCAR assumption does not hold, LD produces biased parameter estimates and biased statistical tests because results are unrepresentative of the population sampled (Cohen & Cohen, 1983). LD, therefore, is not a satisfactory solution to the missing data problem in general (Wilkinson & The APA Task Force on Statistical Inference, 1999).

(2) Pairwise Deletion (PD)

PD retains all available data provided by a subject. If this approach is applied to data analysis, descriptive statistics and a few inferential statistics ($t$, $z$, and chi-square, etc.) are computed from non-missing data on each variable (Glasser, 1964; Raymond & Robert, 1987). It is the default setting in SPSS®, SYSTAT®, and SAS® for descriptive, correlation, and regression analysis using either correlation or covariance matrices. Though PD is as easy as LD, it is not as widely used as LD. Only 7.6% of the studies we reviewed used PD. According to Kim and Curry (1977), PD is an attractive alternative when there is a small number of missing cases on each variable relative to the total sample size, and a large number of variables are involved.

Compared with the LD method, the PD approach utilizes information obtained from partially complete observations. Its disadvantage is that the sample data change from variable to variable. This variability in the sample base creates practical problems, such as the determination of sample size and degrees of freedom. It is especially problematic for multivariate statistical analyses where solutions and intermediate computations are often based on the entire raw data matrix (Rubin, 1987). Cool (2000, p.7) states, “When correlations and other statistics are based on different but overlapping sub-samples of a larger sample, the population to which
generalization is sought is no longer clear. It is possible to compute correlation matrices with mutually inconsistent correlations.” Because of this problem, sample correlation or covariance matrices may not be Gramian (or semi-positive definite) (Malhotra, 1987). Consequently, solutions obtained from factor analysis, structural equation modeling, or other correlation/covariance based modeling or maximum-likelihood based estimation methods are not legitimate (Little & Rubin, 1987). Like LD, the PD method produces biased parameter estimates and biased statistical tests unless the MCAR assumption holds. For these reasons, PD is not a satisfactory solution to the missing data problem (Wilkinson & The APA Task Force on Statistical Inference, 1999).

(3) Mean Substitution (MS)

The MS approach “solves” the missing data problem by replacing the missing values with the mean of the variable (Wilks, 1932). This step is accomplished at the onset of data analysis. It therefore assumes that the mean of the variable is the best estimate for any observation that has missing information on that variable. In contrast to LD and PD, the MS method does not alter the sample mean of the variable and does not discard any information already collected. Although MS is available in some software programs (e.g., SPSS®), only 2% of the studies we surveyed used this method to treat the missing data problem.

A variation of MS is to impute the missing value with a sub-group mean. For example, if the observation with a missing value is a Republican, the mean for all Republicans is computed and inserted in place of the missing value. This procedure is not as conservative as inserting the overall mean of the variable (Tabachnick & Fidell, 2001).

Regardless of which version of MS is applied, this method has many statistical pitfalls. According to Little and Rubin (1987), the limitations include: (a) Sample size is overestimated,
(b) variance is underestimated, (c) correlations are negatively biased, and (d) the distribution of
new values is an incorrect representation of the population values because the shape of the
distribution is distorted by adding values equal to the mean. The bias introduced into the
population variance, correlation, and variable distribution depends on the amount of missing data
and on the actual values that are missing. Little and Rubin’s recommendation is to never use the
MS method.

(4) Simple Hot-Deck (HD)

The HD method replaces each missing value with a randomly drawn value from the set of
data values already collected on the same variable (Reilly, 1993). None of studies in our review
used this method to impute missing data. Parameters estimated by this method have larger
variances, compared to those estimated by MS, but smaller variances than those obtained from
complete data. The most serious drawback of this method is the distortion of correlations and
covariances. Consequently, this method should not be used when statistical methods based on
either correlations or covariances are to be used (see statements earlier regarding the same
problem associated with PD; Little & Rubin, 1987).

(5) Regression Estimation (RE)

The RE method imputes missing values with predicted values derived from a regression
equation based on variables in the data set that contain no missing values. Variables with missing
data are treated as criterion variables and are predicted by all of the variables having complete
data. Four studies reported using this method to handle missing data; they constituted less than
1% of the studies surveyed.

If the missing data exhibit a monotone pattern, as illustrated in Table 2B, and are
assumed to be missing at random—the weaker and more realistic assumption, the RE method
can be used to simplify the estimation of population parameters (see earlier comments referring to this method under “3. Missing Data Mechanism”). Compared to other methods already mentioned, the RE method is more informative because it utilizes information already existing in a data set. The AM procedure in BMDP® applies this method to estimate missing values as well as out-of-range data points (BMDP, 1992, pp. 959-976).

Similar to MS and HD methods, RE has the advantage of preserving cases with missing data, maintaining the sample size. Disadvantages of RE include: (a) a regression model needs to be specified; (b) imputed values are always perfectly predicted from the regression model, thus, correlations and covariances are inevitably inflated; (c) it can be difficult to apply RE to multivariate data sets when more than one variable has missing values; (d) predicted values may exceed the logical range of scores for the missing data; (e) may require large samples to produce stable estimates (Donner, 1982) (f) can produce leptokurtic distributions (Rovine, 1994); (g) if good and relevant predictors of missing data are not available in the data set, predicted values are no better than the mean. In other words, RE and MS yield approximately the same result if an effective regression model cannot be identified.

To overcome limitation (b) discussed above, statisticians have suggested a modified RE method in which imputed values have a random error added to them (Beale & Little, 1975). The random error is randomly generated from a normal distribution with a mean of 0 and a standard derivation equal to the square root of the mean square error of the regression model (Little & Rubin, 1987).

As stated earlier, the MS, HD, and RE ad hoc methods “solve” the missing data problem by imputing missing values once, thus they are referred to as single imputation methods. The single imputation approach unfortunately does not reflect the uncertainty in missing data.
estimates. That is, the error term in the estimation equation (of whatever form) used to impute missing values is set to zero. Furthermore, the sample size is overstated, confidence intervals for estimated parameters are too narrow, and Type I error rates are too high (Little & Rubin, 1987).

To illustrate these deficiencies in the MS, HD, and RE methods, let’s assume that 30% of the data are missing and that we wish to generate a confidence interval about a regression coefficient. The use of any of these three methods to produce a complete data set means that intervals with nominal confidence values of 90%, 95%, and 99% levels in reality have levels of 77%, 85%, and 94%, respectively (Rubin, 1996). Similarly, in a hypothesis-testing framework involving a null hypothesis with ten parameters in which the missing data have been imputed using the MS, HD, or RE methods, the statistical test performed at nominal alpha-levels of 10%, 5%, or 1% is actually conducted at 57%, 45%, and 25% levels, respectively, a problem that worsens as the amount of missing data and/or the number of parameters increases (Rubin).

5. Principled Single Imputation--FIML

FIML stands for Full Information Maximum Likelihood; it represents a principled method for estimating means and covariances based on incomplete data when the missing values are assumed to be missing at random (MAR). In maximum likelihood (ML), parameter estimates are derived such that the likelihood of reproducing the data given the parameter estimates is maximized. The FIML method for estimating parameters in the presence of data missing at random makes extensive use of ML.

FIML has its roots in the work of Hartley and Hocking (1971). Given $q$ groups, one for each pattern of missing data, FIML first calculates the likelihood for each of the $q$ groups. If there are no missing data, then $q = 1$. The intent of FIML is to use the information in each missing data pattern to estimate parameters. Under the assumption of a multivariate normal
population distribution, the $q$ likelihoods are summed. The resulting summed likelihood serves as the basis for finding parameter estimates using ML. FIML is conceptually similar to the $q$-group method of Hartley and Hocking (1971), except that the likelihood is calculated for each case, using whatever data are available for that observation.

To illustrate how FIML estimates parameters in the presence of missing data, let’s suppose that data are collected for $N$ cases on three variables $Y$, $X_1$, and $X_2$, and the researcher’s interest is in estimating the population means, variances, and covariances collected in the vector $\mu$ and the matrix $\Sigma$:

\[
\mu = [\mu_1, \mu_2, \mu_3],
\]

\[
\Sigma = \begin{pmatrix}
\sigma_{11} & \sigma_{12} & \sigma_{13} \\
\sigma_{21} & \sigma_{22} & \sigma_{23} \\
\sigma_{31} & \sigma_{32} & \sigma_{33}
\end{pmatrix}
\]

Suppose, however, that $X_2$ is missing for some cases. Under the assumption of a multivariate normal distribution and MAR, FIML computes a likelihood for each case using whatever data are available. If case #1 has no data for $X_2$, it contributes only to the estimation of $\mu_1$, $\mu_3$, $\sigma_{11}$, $\sigma_{13}$, $\sigma_{31}$, and $\sigma_{33}$. After $N$ likelihoods have been computed, they are summed and ML is used to estimate the means, variances, and covariances based on the summed likelihood. These parameter estimates will be unbiased and efficient in the presence of missing data under MAR and multivariate normality assumptions.

**Illustration of FIML using the AMOS Software**

In this section, we illustrate how different methods of handling missing data, FIML versus LD, can yield different parameter estimates and, potentially, different inferences. To
illustrate these effects, we used a data set from 1,302 U.S. colleges and universities for the 1993-94 school year.

Data Set

The data set was provided by US News & World Report and is available from http://portfolio.iu.edu/peng/articles/persist.sav. For this data set, hereafter referred to as the Persist data, we were interested in modeling variability in student persistence, as measured by graduation rates, as a function of college-related predictors, such as tuition and institutional quality. Because the cases were colleges and universities, each institution’s graduation rate (GRADRATE), in percentages, was the unit of analysis and served as a proxy for student persistence.

Three (latent) variables, often described in the literature as predictors of student persistence, appeared in the path model: ability, quality of institutions, and costs. These were manifested in financial cost (room and board or RMBRD), additional student fees (ADDFEES), in-state tuition (INSTATE), cost of books (BOOKCOST), selectivity (or quality) of the institution (SELECT), percent of faculty with Ph.D.'s (PCTPHD), student to faculty ratio (STUDFACT), percent of full time students (PCTFULL), expenditure per student (PERSPEND), and student quality and preparedness, measured by the average SAT Math and Verbal scores (AVESAT). The amount of missing data for these variables ranged from 0.2% to 40.3%, with the percentage of missing data for graduation rate equal to 7.5%. All variables thought to contribute to the missingness were included in FIML, even if some of them were not included in subsequent analyses (Schafer, 1997).

Statistical Modeling
The AMOS software (Arbuckle, 1995) was used to fit the path model in Figure 1 to the Persist data. Both FIML and LD methods were used to handle missing data. For each method, we examined its effect on model conclusions based on estimated path coefficients and their standard errors. We did not attempt to find the best-fitting model under each method for two reasons: first, the best model might be different and second, different best models might be attributable to the vagaries of finding these models as well as to some intrinsic properties of the missing data methods. Because of these reasons, the path model was fit once by each method to the data. Table 4 presents the estimates and standard errors of the path coefficients using both FIML and LD.

**Comparison of AMOS FIML Results with the LD Results**

The results in Table 4 show several differences in the estimated path coefficients and their standard errors as a function of the missing data method used. FIML produced larger estimated path coefficients for eight out of twelve paths. The size of the standard errors for the two methods was equally divided, with FIML standard errors larger than LD-based values in six out of twelve paths. In two paths, the signs reversed between the FIML- and LD-estimated coefficients. For example, the estimated path coefficient of QUALITY on student’s graduation rate (GRADRATE) using FIML was $-3.48$, whereas the LD-based coefficient was $3.43$. The FIML coefficient ($-3.48$) means that an increase in institutional quality was associated with a decrease in student graduation rates, with other effects (or paths) held constant. The LD coefficient ($3.43$) suggested that increasing institutional quality was associated with higher student graduation rates. These results provide empirical evidence that the choice of method for handling missing data can substantially affect inferences.

**Comments on FIML**
Though FIML has been implemented in the AMOS software (Arbuckle, 1999), the study of FIML has been primarily in structural equation modeling. Arbuckle (1996), Enders and Bandelos (1999), and Wothke (2000) examined FIML for various structural equation models, sample sizes, and the percent of missing data. Results from these studies indicate that, as expected, parameters estimated using FIML are unbiased and efficient when the data are multivariate-normal and the MAR assumption holds. However, failure to satisfy the multivariate normality assumption can dramatically bias FIML estimates.

Another drawback of FIML is that tests of goodness-of-fit are not always available. For example, in regression modeling with missing data on the dependent variable \( Y \), the estimates of regression coefficients are unbiased and efficient under the assumption of MAR and normality. However, a test of the overall model fit to data is not possible because there is no value of \( N \) applicable to the entire sample. Similar problems with FIML emerge in structural equation modeling in which the goodness-of-fit test and several commonly prescribed measures of fit are not available. Another potential problem is that the covariance matrix may be indefinite, which can lead to significant parameter estimation difficulties, although these problems are often modest (Wothke, 2000).

A practical issue with using FIML in the presence of missing data is that variables believed to be predictive of missingness must be in the regression or structural equation model, even when these variables are not of substantive interest. In some instances, the presence of such variables may have an important impact on variables of substantive interest in the model.

**6. Principled Single Imputation—SPSS® EM**
EM stands for Expectation-Maximization; it is a principled method for handling missing data. Dempster, Laird, & Rubin in their seminal paper on this method coined the term EM (1977). According to Little and Rubin (1987),

The EM algorithm formalizes a relatively old ad hoc idea for handling missing data: (1) replace missing values by estimated values, (2) estimate parameters, (3) re-estimate the missing values assuming the new parameter estimates are correct, (4) re-estimate parameters, and so forth, iterating until convergence (p. 129).

Each iteration of the EM algorithm consists of two steps: an E (Expectation) step followed by an M (Maximization) step. In the E step, the expectation of the complete data log-likelihood is derived, given the observed data and the estimated parameters from a previous iteration. In the M step, the conditional expectation of the complete data log-likelihood is maximized. The observed data log-likelihood is increased until a stationary point is reached (Dempster et al., 1977). In other words, the algorithm continues until the observed likelihoods produced in two consecutive iterations are almost identical. Only four studies (less than 1%) we reviewed used the EM method to treat missing data problem.

The Missing Value Analysis (MVA) module available in SPSS® version 10 and beyond implements the EM algorithm indirectly as a single imputation method. First, it employs the EM algorithm to derive ML estimates of parameters based on a researcher’s specification of a probability model for the data (e.g., multivariate-normal). Operationally, this involves using a series of regressions for each missing data pattern similar to Buck’s RE method (Buck, 1960).

After the EM algorithm converges, MVA computes imputed values for missing data based on means and covariances estimated from the last iteration of the EM algorithm. The imputed values replace missing data to yield a complete data set, which is subsequently analyzed
by the statistical method(s) of a researcher’s choice. We consider the EM algorithm used in MVA to be a principled, single-imputation method for handling missing data.

**Illustration of SPSS® EM Method**

In this section, we describe how the SPSS® EM Method was applied to impute values for missing data for a multivariate data set. This was followed by fitting a logistic regression model in order to help predict adolescent’s behavioral risk. Results are presented in terms of regression coefficient estimates and standard errors.

**Data Set**

Self-reported health behavior data were collected from 517 adolescents enrolled in two junior high schools (grades 7 through 9) in the fall of 1988. On the day of data collection, two questionnaires were administered: the Health Behavior Questionnaire (HBQ; Ingersoll & Orr, 1989; Resnick, Harris, & Blum, 1993) and Rosenberg’s self-esteem inventory (Rosenberg, 1965). Among the 517 students, 85 did not complete all questions. Thus, the sample size with complete information was 432 (83.4% were Whites and the remaining Blacks or others), with a mean age of 13.9 years and nearly even numbers of girls (n=208) and boys (n=224).

The HBQ asked adolescents to indicate whether they engaged in specific risky health behaviors (Behavioral Risk Scale) or had experienced selected emotions (Emotional Risk Scale). The response scale ranged from 1 (*never*) to 4 (*about once a week*) for both scales. Cronbach’s alpha reliability was 0.84 for the Behavioral Risk Scale and 0.81 for the Emotional Risk Scale (Peng & Nichols, in press). Adolescents’ self-esteem was assessed using Rosenberg’s self-esteem inventory (Rosenberg, 1965). Self-esteem scores ranged from 9.79 to 73.87 with a mean of 49.97 and standard deviation of 10.09. Furthermore, among the 432 adolescents, 12.27% (or 53) indicated an intention to drop out of school; 44.68% (or 193) were from intact families,
22.69% (or 98) were from families with one step-parent, and 32.63% (or 141) were from families headed by a single parent. The data set is hereafter referred to as the Adolescent data [available from http://portfolio.iu.edu/peng/articles/logregdata(peng).sav as a SPSS® data file or from http://portfolio.iu.edu/peng/articles/logregdata.por as a portable data file].

Research Question and Statistical Modeling

For the Adolescent data, we were interested in identifying adolescents at the greatest behavioral risk from their gender, intention to drop out from school, family characteristics, emotional risks, and self-esteem scores. Given this objective, the research hypothesis was stated as follows: “The likelihood that an adolescent is at high behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self-esteem.” Scores on the Behavioral Risk Scale of the HBQ ranged from 40.44 to 95.21 with a mean of 47.23 and a standard deviation of 8.40, and it was decided that adolescents scoring 52 or above (i.e., more than a half of SD above the mean) were identified as being at high behavioral risk, and cases with scores below 52 not at high risk.

A logistic regression model was fit to the data using adolescents’ risk level on the Behavioral Risk Scale of the HBQ as the dependent variable and gender, intention to drop out of school, type of family structure, emotional risk, and self-esteem score as the explanatory variables.

SPSS® EM Method for Imputing Missing Data

As stated earlier, out of the 517 students, 85 did not complete all questions. Most missing data occurred on Behavioral Risk Scale (77 cases) or Emotional Risk Scale (34 cases); only six adolescents did not indicate if they intended to drop out from school. To impute values for missing data, the entire data set was submitted to the MVA module in SPSS® version 11.01. The
imputation model specified included gender, intention to drop from school, behavioral risk score, emotional risk score, self-esteem score, family configuration (intact family, step family, or single parent family), and living arrangement (living with both parents, father alone, mother alone, neither, father and step-mother, or mother and step-father). These variables were believed to be predictive of the missingness for the Adolescent data.

Once the imputation model was defined in MVA and the EM estimation method was selected, we further specified a multivariate normal distribution for all variables selected and 50 iterations for the EM iteration. The result was a complete data with missing data replaced by imputed values at the last iteration of the EM algorithm. This complete data set was subsequently analyzed by the logistic regression procedure in SAS® version 8.2.

Analysis Results

Using the criterion of 52 points on the Behavioral Risk Scale of the HBQ, 102 adolescents were identified to be at high behavioral risk while 415 were not. To entertain the research hypothesis that “the likelihood that an adolescent is at high behavioral risk is related to his/her gender, intention to drop out of school, family structure, emotional risk, and self-esteem,” a logistic regression model was fit to the data to yield the following result:

\[ \text{Predicted logit (} Y=\text{high behavioral RISK}) = -1.6120 + (1.1014) GENDER + (2.5369) DROPOUT + (0.4084) FAMILY + (0.00809) EMOTION + (-0.0443) \text{ ESTEEM.} \tag{5} \]

According to this model, the log of the odds of an adolescent being at high behavioral risk was positively related to gender \((p<.0001, \text{ Table 5})\), intention to drop out of school
(p<.0001), and family structure (p<.01); it was negatively related to self-esteem (p<.01), and insignificantly related to emotional risk (p=0.5405). The effectiveness of Eq. (5) in explaining the data was examined using (a) overall model evaluations, (b) statistical tests of each explanatory variable, (c) goodness-of-fit statistics, and (d) measures of association. These indices were recommended in the literature for evaluating logistic regression results (Peng, Lee, & Ingersoll, 2002; Peng, So, et al., 2002); and they are presented in Table 5.

Comparison of SPSS® EM Results with the LD Results

The same logistic regression model was fitted to 432 cases with complete data. This was equivalent to treating the missing data with the listwise deletion. Results are presented in Table 6. According to Table 6, the LD approach led to the following logistic model:

\[
\text{Predicted logit (Y=high behavioral RISK)} = -1.6341 + (1.0470)\text{*GENDER} + (2.0772)\text{*DROPOUT} + (0.4628)\text{*FAMILY} + (0.0065)\text{*EMOTION} + (-0.040)\text{*ESTEEM}.
\] (6)

According to Eq. (6), the log of the odds of an adolescent being at high behavioral risk was positively related to gender (p<.0002, Table 6), intention to drop out of school (p<.0001), and family structure (p<.003); it was negatively related to self-esteem (p<.02), and was not related to emotional risk (p=.6459). Essentially, Eq. (6) uncovered the same relationship as Eq. (5) did between the likelihood of high behavioral risk and the five explanatory variables. However, there were important differences between these two results. First, standard errors of the estimated regression coefficients were in general larger in Table 6, compared to those shown in Table 5, except for EMOTION and ESTEEM. Second, the significance levels associated with Wald’s chi-
square tests were larger in Table 6 than in Table 5, most noticeably on variable ESTEEM. Finally, measures of association shown at the bottom of Table 6 were smaller than those of Table 5. These differences speak to the consequences of decreasing sample size and lowering statistical power that resulted from applying the LD method.

**Comments on EM and SPSS® EM Method**

Applications of EM-type algorithms have a long history (McLachlan & Krishnan, 1997, pp. 34-37). The popularity of the EM algorithm among statisticians is largely based on the fact that this approach allows many complex statistical problems to be reformulated as a missing data problem in a way that greatly simplifies parameter estimation (e.g., mixture models, random effects models, hierarchical linear models, unbalanced designs including repeated measures). In most applications, it is assumed that data follow a multivariate normal distribution.

In addition to taking the missing data mechanism into account, the primary advantages of the EM algorithm are simplicity and ease of computing (Dempster, et al. 1977; Little & Rubin, 1987). Since fewer than 1% of the studies we reviewed reported using the EM method to handle missing data, we suspected that the low frequency of usage might be attributable partly to its drawbacks. First, the M-step of the algorithm does not produce sampling variances of the parameter estimates. Several extensions of the EM algorithm have appeared in the statistical literature that estimate standard errors (Meng & Rubin, 1991; McLachlan & Krishnan, 1997, pp. 28-29); but in general these methods have not been implemented in missing data programs such as the MVA in SPSS®. Second, all likelihood-based methods including the EM method are iterative and model-specific (or model-dependent). Because of their dependency on specific models, they are not yet programmed into all statistical analysis procedures in a general-purpose statistical package, such as SPSS® regression or ANOVA. Third, the convergence rate during
iterations is proportional to the percent of completely observed data in a data set (Little & Rubin, 1987, p.130). As a result, the rate of convergence can be painfully slow if the percentage of missing data is high (Buu, 1999; Little & Rubin, 1987).

Buu (1999) investigated the Type I error rate control and the precision of parameter estimation of three missing data methods: the SPSS® EM approach (i.e., MVA), the direct, one-step EM algorithm, and SAS PROC MIXED. Her study revealed that the SPSS MVA approach rejected the null hypothesis more frequently than it should, resulting in inflated Type I error rates, though it estimated parameters accurately. These findings were obtained in repeated-measures ANOVA using both empirical data and data simulated from multivariate normal distributions. The direct, one-step EM and SAS PROC MIXED yielded very similar and correct Type I error rates with SAS PROC MIXED slightly outperforming the EM algorithm. Buu’s study is unique in the literature as most missing data methods were developed and evaluated in between-subjects designs rather than within-subjects or time-series studies (Roth, 1994).

7. Multiple Imputation

To overcome the limitations of methods that fail to take into account the uncertainty associated with imputed values, Rubin and his associates developed the multiple imputation (MI) method in the 1980’s. MI is a valid method for handling missing data under the MAR condition. It is a general-purpose method, highly efficient even for small sample sizes (Graham & Schafer, 1999). None of the articles we reviewed used this method. For an introduction to MI, we recommend Schafer’s 1999 paper (Schafer, 1999). MI generally consists of three steps: imputation, analysis, and pooling. In the first (imputation) step, each missing value is replaced by not one, but $m > 1$ simulated values. Imputed values are drawn from a distribution that is specified by the researcher. At the end of the imputation step, $m$ complete data sets are created.
In the second (analysis) step, each of the $m$ complete data sets is analyzed by standard complete-data methods. Finally, results of the $m$ analyses are integrated in the third (pooling) step to yield a final result such as an interval estimation of a population parameter, a $p$-value of null-hypothesis testing, or a likelihood-ratio test statistic. In the next section, we elaborate on each of these three steps applying PROC MI and PROC MIANALYZE in SAS® version 8.2 (SAS Institute., 1999) to a real-world data set.

**Illustration of the MI Method**

The Adolescent data set was once again used to illustrate the MI method. The research question posed to these data and the logistic modeling approach are described in Section 6 under the heading, Research Question and Statistical Modeling.

**Step 1—Imputation**

In this step, each missing value is replaced by $m>1$ imputed values. The model we adopted for imputation was the same logistic regression model that would be used subsequently to analyze the data, though these two models need not be identical (see the section, Comments on the MI Method and Schafer, 1997, pp. 139-143). The data matrix consisting of one dependent variable (the behavioral risk score) and five explanatory variables was submitted to PROC MI in SAS® version 8.2 to impute missing data 5 times (the default). Since the missing rate for the current data was 16.44%, $m=5$ was considered sufficient (Graham & Schafer, 1999). The entire SAS® program is available from [http://portfolio.iu.edu/peng/articles/lr-em.fit.sas](http://portfolio.iu.edu/peng/articles/lr-em.fit.sas); the syntax for PROC MI was as follows:

```sas
PROC MI DATA=risk SEED=37851 OUT=outmirisk ROUND=1 1 1 1 1 1; VAR behrisk gender dropout emotion esteem family;
```
The upper case commend is SAS® keywords and lower cases are user-specified variable names or SAS® data set names. The keyword, SEED, specifies a starting value for the random number generator whereas ROUND specifies units of rounding for imputed values. To revalidate the imputation results, a researcher can reset SEED to a different number that will result in a different set of initial estimates for parameters (i.e. regression coefficients for the Adolescent data set).

PROC MI assumes that missing data were missing at random (the MAR condition) and that the population data distribution is multivariate-normal. Both assumptions were deemed reasonable for the present data (see Section 8 on the importance of satisfying these two assumptions). At the end of the imputation, 5 data sets with $N=517$ cases were created. Table 7 summarizes the increase in variability as a result of multiple imputations for the three variables that had missing data, i.e., dropout, emotional risk, and behavioral risk. The increase in variability reflected the uncertainty in the imputed values—an uncertainty not accounted for by single imputation methods. Thus, variables (e.g., behavioral risk) associated with larger amount of missing data tend to have greater uncertainty (i.e., increase in variation) in estimated parameters than those associated with smaller amounts of missing data (e.g., dropout and emotion risk).

Step 2—Analysis

In the second (analysis) step, each of the $m$ complete data sets is analyzed by a standard complete-data method, such as the logistic model prescribed for the Adolescent data. Consequently, each data set obtained from Step 1 above was analyzed by PROC LOGISTIC in SAS® version 8.2 to determine if and how the likelihood that an adolescent was at high behavioral risk was related to his/her gender, intention to drop out of school, family structure,
emotional risk, and self-esteem. The SAS® syntax for PROC LOGISTIC for these five data sets was as follows:

```sas
PROC LOGISTIC DATA=outmirisk;
   MODEL risk=gender dropout family emotion esteem/
       RSQUARE LACKFIT;
   FORMAT risk br.;
   BY _imputation_;
```

Again, the upper case commend is a SAS® keyword and lower cases are user-specified variable names or SAS® data set names. The statement, FORMAT, applied labels (such as “high” and “low”) to numerical codes (such as “1” and “2”) of the dependent variable, risk. The BY statement applied the logistic regression modeling repeatedly to all imputed data sets. The variable name, _imputation_, was created in the preceding PROC MI to denote each imputed (complete) data set. At the end of Step 2, five logistic regression results, similar to Tables 5 or 6, were obtained.

**Step 3—Pooling**

The last stage in MI is to pool results of the $m$ analyses to yield a final result, using a formula available in Rubin (1987) and Schafer (1997). To accomplish this step, we invoked PROC MIANALYZE in SAS® 8.2 using the SAS® syntax as follows: [the entire SAS® program is available from http://portfolio.iu.edu/peng/articles/mianalyze-lr.fit(5var).sas]

```sas
PROC MIANALYZE DATA=outest;
   VAR intercept sex drop famstruc emorisk selfest;
RUN;
```

In the program above, the keyword DATA= specifies a covariance matrix of parameter estimates (i.e., the intercept and logistic regression coefficients) obtained from PROC
LOGISTIC in Step 2 above. From this covariance matrix, PROC MIANALYZE was able to yield a single (pooled) estimate of six logistic regression estimates (Table 8).

According to Table 8, the log of the odds of an adolescent being at high behavioral risk was linearly related to five explanatory variables in the following pattern:

\[
\text{Predicted logit (Y=high behavioral RISK) } = -1.9398 + (1.1860) \times \text{GENDER } + \\
(2.0751) \times \text{DROPOUT } + (0.4343) \times \text{FAMILY } + (0.0090) \times \text{EMOTION } + \\
(-0.035) \times \text{ESTEEM.} \quad (7)
\]

The interpretation of the significance of the overall model and each explanatory variable was very similar to that of Eq. (5). In other words, the predicted logit was positively related to gender \((p<.002)\), intention to drop out of school \((p<.0001)\), and family structure \((p<.003)\); it was negatively related to self-esteem \((p<.02)\), and insignificantly related to emotional risk \((p=.4884)\).

The comparison of this model to Eq. (6), based on the LD method, is presented in the next section.

The pooled estimates shown in Table 8 are computed using formulae from Rubin (1987). For an imputed data set \(i\), let the regression coefficient estimate for an explanatory variable be denoted as \(b_i\) and its variance be \(V_i\), \(i=1 \ldots, m\) (=5 in the present illustration). The final, pooled point estimate for the regression coefficient is the average of all \(b_i\)'s:

\[
\bar{b} = \frac{1}{m} \sum_{i=1}^{m} b_i. \quad (8)
\]

The within-imputation variance of these point estimates is given by

\[
\text{WV} = \frac{1}{m} \sum_{i=1}^{m} V_i. \quad (9)
\]

The between-imputation variance of these point estimates is given by
BV = \frac{1}{m-1} \sum_{i=1}^{m} (b_i - \bar{b})^2. \quad (10)

Thus, the total variance associated with the pooled estimate \(\bar{b}\) is

\[ TV = WV + \left(1 + \frac{1}{m}\right)BV. \quad (11) \]

From the total variance, Rubin (1987) derived the sampling distribution (i.e., \(t\)) for the statistic \((b - \bar{b}) / \sqrt{TV}\) with the \(df_m\) equal to

\[ df_m = (m - 1)[1 + \frac{WV}{(1 + m^{-1})BV}]^2. \quad (12) \]

The above \(df_m\) and its associated \(t\)-statistic were used by PROC MIANALYZE to compute the significance level \((p)\) and 95% confidence intervals around estimated regression coefficients, as reported in Table 8. The level of confidence interval may be altered by researchers to a level other than 95%. Pooled estimates can also be computed by PROC MIANALYZE for a variety of statistical indices such as, means, variances, correlations, \(p\)-values, odds ratios, likelihood-ratio test statistic, and covariance matrices in both univariate and multivariate cases (Rubin, 1987; SAS Institute, 2000; Schafer, 1997).

**Comparison of MI Results with the LD Result**

Comparing results presented in Eq. (7) and Table 8 to those in Eq. (6) and Table 6, one notices that both models uncovered the same relationship between the logit of the dependent variable and the five explanatory variables. There were, however, differences between these two results. First, standard errors were generally larger in Table 6 than in Table 8, except for the CONSTANT (i.e., the intercept) and the GENDER variable. Yet, the significance levels associated with the \(t\)-test in Table 8 were larger for variables GENDER, FAMILY, and ESTEEM than those associated with Wald’s chi-square tests in Table 6. This difference illustrates how MI
captures the uncertainty associated with missing values and, consequently, leads to more valid statistical inferences in terms of both null hypothesis testing and interval estimation of the regression coefficients.

**Comments on the MI Method and PROC MI in SAS®**

Key to the success of the MI method is the specification of the imputation model. MI assumes that the imputation model is identical to the model that a researcher will use to analyze the data (i.e., the analysis model). In practice, though, these two models need not be identical. According to Schafer (1997), the imputation model should contain predictors that are important substantively, highly predictive of missingness and of variables with missing data, and reflect special features of the sample design (e.g., probability surveys). The imputation model does not have to be conceptually meaningful. In practice, it typically contains the variables of substantive interest as well as predictors believed to reflect missingness. There is empirical evidence that the imputation model can be useful in reducing bias even if it is mis-specified (Schafer).

MI is a Bayesian approach; it is similar to the EM method in that maximum likelihood estimates of population parameters are calculated based on observed data only. Both approaches are iterative and must converge to a criterion established by the programmer or the researcher. Both methods summarize a likelihood function that has been integrated over a conditional distribution for the missing data, conditional on the observed data and estimated parameters. The major difference between these two is that EM accomplishes this task by numerical algorithms whereas the MI method does so by Monte Carlo approaches. Specifically, PROC MI in SAS® and the NORM program written by Schafer use a data augmentation algorithm to create a Markov chain from which random draws of missing data are sampled from the conditional distribution of missing values, conditional on the observed data and estimated parameters (SAS...
Advances in Missing Data 37

Institute, 2000; Schafer, 1997). These parameter estimates are themselves random draws from the posterior distribution finalized after the Markov chain becomes stationary.

The data augmentation algorithm used by PROC MI assumes multivariate normality for the complete data in order to impute values for missing data, or just enough missing values so that the imputed data sets have monotone missing patterns (SAS Institute, 2000). Violation of the multivariate normality assumption had little impact on inferences regarding the first and the second moments of the population distribution, i.e., the mean and the variance, as reported in Graham and Schafer (1999).

The efficiency of a parameter estimate based on $m$ imputations is $(1 + \gamma/m)^{-1}$, where $\gamma$ is the rate of missing information (Rubin, 1987, p.114). Table 9 relates the rate of missing ($\gamma$) with the number of imputations ($m$) for efficiency of recovery of the true parameter. It is clear from Table 9 that, for low rates of missing data (20% or less), no more than $m= 3$ imputations are needed to be at least 90% efficient. For higher rates of missing, $m= 5$ or 10 imputations are needed. In general, it is recommended to impute missing values 5 times before results are pooled (Schafer & Olsen, 1998). The MI procedure in SAS® defaults the number of imputations to 5.

Despite its strengths, there are several concerns with MI that have been documented in the literature. The first is that each application of MI produces slightly different imputed values and associated statistics. Consequently, results cannot always be replicated if MI is used to handle missing data. Other things being equal, this issue increases in importance as the amount of missing data increases. Second, Allison (2001) discusses difficulties with MI when interest is in estimating interaction effects. A third concern is the limited availability of MI software in widely used statistical software. In addition to the two procedures (MI and MIANALYZE) implemented in SAS® version 8.2 and beyond, a free MI software NORM, written by Schafer
Advances in Missing Data (1997), can also be utilized to perform the multiple imputation under the multivariate normal assumption. This software is available from http://www.stat.psu.edu/~jls/misoftwa.html, along with three additional modules: CAT, MIX, and PAN. CAT performs MI for categorical data under the loglinear distribution assumption, MIX performs MI for a mixture of continuous and categorical data under the general location distribution assumption, and PAN performs MI for panel data or clustered data under a multivariate linear mixed-effects distribution. Although Schafer’s software allows MI to be applied to a variety of missing data problems, some educational researchers may find these programs to be less than user-friendly.

8. Importance of Statistical Assumptions for Principled Methods

The FIML, EM, and MI methods described in the preceding sections rely on two key assumptions. One is that the missing data are missing at random. To maximize the likelihood of satisfying the MAR assumption, researchers need to plan for ignoralibility in the design of a study (Heyting & Tolboom, 1994). In other words, if a researcher anticipates the occurrence of missing data, he/she needs to plan to collect data for variables believed to be related to, or predictive of, missingness as a routine part of the design of a study, even though these variables may be peripheral to the research question under pursuit.

According to Graham and Schafer (1999), data-analytic results are relatively insensitive to the misspecification of the missing data model because such a model does not distort the entire data distribution, only the portion that is missing. Similarly, Schafer (1997) argued that a missing data model that explains some, but not all, of the missingness reduces the bias accordingly, and is likely to produce parameter estimates superior to those based on LD. In a small simulation study investigating the consequences of failing to satisfy the MAR assumption, Schafer showed that treating the missing data as though they were missing at random produced less biased means.
and standard deviations than LD. Ezzati-Rice, Johnson, Khare, Little, Rubin, and Schafer (1995) and Heitjan and Basu (1996) reported similar findings. In sum, the best advice is for researchers to plan on collecting data so as to maximize the likelihood of satisfying the ignorability condition.

The FIML, EM, and MI methods (as implemented in SAS® PROC MI and the stand-alone NORM program) also assume that complete data follow a multivariate-normal distribution. Again, there is surprisingly little literature examining effects of nonnormality on parameter estimates. While it may seem reasonable to conclude that drastic departures from this assumption will seriously bias inferences derived under these methods, some authors disagree. Graham and Schafer (1999) claimed that nonnormality had little impact on inferences derived from MI because the imputed values would still resemble the observed data in their first and second moments (i.e., the mean and the variance). Since most data analyses of interest in educational research are based on the first and second moments, it can be argued that the damage inflicted by nonnormality is usually not severe enough to significantly bias inferences. Graham and Schafer went on to argue that nonnormality is more of a threat to inferences based on parameters other than the first or second moments, for example, a variable’s 95th percentile. Despite the optimism of Graham and Schafer, it seems prudent to use missing data methods cautiously when there is evidence of substantial nonnormality. And there is some evidence that parameter estimates under FIML are not robust to departures from multivariate normality (Arbuckle, 1996; Enders, 2001).

9. Implications and Recommendations for Educational Research

Just as qualitative researchers in education have struggled to keep up with rapidly emerging approaches during the past 20 years, so must quantitative researchers incorporate advances provided by methodologists in their realm of inquiry. The impact of missing data on quantitative research should be a concern to educational researchers. It has been addressed in the
methodological literature (e.g. Afifi & Elashoff, 1966; Becker & Powers, 2001; Becker & Walstad, 1990; Cool, 2000; Coons, 1957; Kim & Curry, 1977; Rubin, 1987; Schafer, 1999; Tirri & Silander, 1998), and substantial progress has been made in the last two decades of the 20th century (Little & Rubin, 1987; Schafer, 1997; special section of Psychological Methods, December 2001).

However, as shown in our review of quantitative studies published in 11 education journals between 1998 and the summer of 2002, this progress has not had an impact on the way educational researchers handled missing data. In studies where missing data were reported or detected, 90% treated the problem with either the LD or the PD method, 3% did nothing about this problem, and less than 1% used the principled EM method. Unfortunately, both the LD and the PD methods have been shown in the literature to be biased and inefficient. An APA task force (Wilkinson & Task Force, 1999) warned against their use in empirical research.

Newer and more principled methods take into account conditions under which missing data occur. These methods have been implemented into statistical software (e.g., SPSS® and SAS®); thus, they are accessible to educational researchers. In this paper, we demonstrated the use of three of these principled methods for treating missing data, namely, the Full Information Maximum Likelihood (FIML) method, the SPSS® Expectation-Maximization (EM) method, and the Multiple Imputation (MI) method. These methods were illustrated with real-world data sets in which information was missing from continuous variables in a few cases. Results were interpreted within the framework of assumptions made about missing data. It was shown through these illustrations that different outcomes could be obtained from different treatments of missing data. To minimize, if not eliminate, potential bias in quantitative research findings and to ensure
credible conclusions upon which we base much of our thinking, it is essential that educational researchers keep the following recommendations in mind.

First and the foremost, efforts should be exerted to collect data to the fullest extent and of the highest quality. By so doing, educational researchers keep missing data to a minimum and, therefore, reduce bias and distortion in estimating population parameters or testing pertinent hypotheses. Even though principled procedures for handling missing data have been devised, their validity depends on missing data mechanism and proper specifications of models for treating missing data. Everitt (1998, p.9) asserted that “the percentage of missing data in a study can be considered as one indicator of the quality of the data and, therefore, the quality of the study. It is important not to be seduced into thinking that investigations that are carried out poorly can be rescued by sophisticated statistical analysis.” Since no one procedure has proven to be uniformly superior to others, we concur with Anderson et al. (1983) that “…the only real cure for missing data is to not have any” (p.480).

Second, educational researchers should provide sufficient and consistent information in research reports so that readers may be able to evaluate the soundness of findings based on the methodology. Specifically, educational researchers should always report the actual sample size used in the analysis, the presence or absence of missing data, the cause(s) of missing data, and the treatment of missing data. If results are reported in percentages, it is necessary to also include the sample size on which the percentage is based. Summary statistics such as the sample size, the number of non-responses, the response rate, and $df$ of $t$, chi-square, or $F$ statistics should be presented clearly and consistently in the text and/or summary tables throughout the report.

Third, if missing data are present, researchers must determine whether the cause and pattern of the missing data will seriously impair the quality of the inferences derived and which
procedure, if any, is most appropriate for handling missing data. A careful examination of factors causing missing data and the missing data pattern allows researchers to decide if and how to best deal with missing data in a study. To this end, Table 3 may be a good place to start.

Lastly, it is our hope that these recommendations will move educational researchers as well as journal referees and editors toward formulating new standards in research practices and editorial policies regarding the treatment of missing data in quantitative research. These new standards should enhance the quality of research results and of publications in education. The improved quality of research findings should in turn help to construct a knowledge base that is supported by sound methodologies.
References


Table 1

*Summary of Missing Data Methods Used in Research Published in Selected Education Journals*

<table>
<thead>
<tr>
<th>Journal</th>
<th>Duration</th>
<th># of Studies, N / Articles</th>
<th>Studies with or without missing data</th>
<th>Missing data method used</th>
<th>Complete</th>
<th>Missing, n</th>
<th>Unknown</th>
<th>LD</th>
<th>PD</th>
<th>MS</th>
<th>HD</th>
<th>RE</th>
<th>EM</th>
<th>MI</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>AERJ</td>
<td>1998-1999</td>
<td>29/27</td>
<td>2(7%)</td>
<td>6(21%)</td>
<td>1(5%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(5%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>42/41</td>
<td>6(14%)</td>
<td>31(74%)</td>
<td>5(12%)</td>
<td>24(77%)</td>
<td>3(10%)</td>
<td>2(6.5%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2(6.5%)</td>
<td></td>
</tr>
<tr>
<td>ER</td>
<td>1998-1999</td>
<td>3/3</td>
<td>0</td>
<td>2(67%)</td>
<td>1(33%)</td>
<td>2(100%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>1/1</td>
<td>0</td>
<td>1(100%)</td>
<td>0</td>
<td>1(100%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JCP</td>
<td>1998-1999</td>
<td>108/73</td>
<td>40(37%)</td>
<td>41(38%)</td>
<td>27(25%)</td>
<td>35(85%)</td>
<td>5(12%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(3%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>117/97</td>
<td>30(26%)</td>
<td>73(62%)</td>
<td>14(12%)</td>
<td>58(80%)</td>
<td>11(15%)</td>
<td>3(4%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>JEP</td>
<td>1998-1999</td>
<td>147/111</td>
<td>47(32%)</td>
<td>71(48%)</td>
<td>29(20%)</td>
<td>64(90%)</td>
<td>7(10%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>221/160</td>
<td>82(37%)</td>
<td>116(53%)</td>
<td>23(10%)</td>
<td>109(94%)</td>
<td>5(4%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2(2%)</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>JRST</td>
<td>1998-1999</td>
<td>34/33</td>
<td>11(32%)</td>
<td>18(53%)</td>
<td>5(15%)</td>
<td>16(89%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2(11%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>36/35</td>
<td>10(28%)</td>
<td>23(64%)</td>
<td>3(8%)</td>
<td>16(70%)</td>
<td>2(9%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(4%)</td>
<td>0</td>
<td>4(17%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>JSE</td>
<td>1998-1999</td>
<td>21/21</td>
<td>3(14%)</td>
<td>15(72%)</td>
<td>3(14%)</td>
<td>13(87%)</td>
<td>1(7%)</td>
<td>1(7%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>25/25</td>
<td>6(24%)</td>
<td>16(64%)</td>
<td>3(12%)</td>
<td>13(81%)</td>
<td>2(13%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(6%)</td>
<td></td>
</tr>
<tr>
<td>JSP</td>
<td>1998-1999</td>
<td>29/28</td>
<td>4(14%)</td>
<td>16(55%)</td>
<td>9(31%)</td>
<td>15(94%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(6%)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>49/47</td>
<td>13(27%)</td>
<td>32(65%)</td>
<td>4(8%)</td>
<td>26(84%)</td>
<td>2(6%)</td>
<td>0</td>
<td>0</td>
<td>1(3%)</td>
<td>1(3%)</td>
<td>0</td>
<td>1(3%)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLJ</td>
<td>1998-1999</td>
<td>30/30</td>
<td>13(43%)</td>
<td>13(43%)</td>
<td>4(13%)</td>
<td>12(92%)</td>
<td>0</td>
<td>1(8%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>2000-2002</td>
<td>23/21</td>
<td>8(35%)</td>
<td>11(48%)</td>
<td>4(17%)</td>
<td>9(82%)</td>
<td>1(9%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1(9%)</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>-----------</td>
<td>-------</td>
<td>------------</td>
<td>-----------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td>-------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RHE</td>
<td>56/54</td>
<td>66/66</td>
<td>7(13%)</td>
<td>28(50%)</td>
<td>4(6%)</td>
<td>41(62%)</td>
<td>21(32%)</td>
<td>36(88%)</td>
<td>2(5%)</td>
<td>1(2%)</td>
<td>2(5%)</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>23/19</td>
<td>18/17</td>
<td>10(44%)</td>
<td>6(26%)</td>
<td>8(44%)</td>
<td>7(39%)</td>
<td>6(100%)</td>
<td>6(86%)</td>
<td>1(14%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TRSE</td>
<td>4/4</td>
<td>5/5</td>
<td>1(25%)</td>
<td>2(50%)</td>
<td>3(60%)</td>
<td>1(25%)</td>
<td>2(100%)</td>
<td>3(100%)</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1087/918</td>
<td>305(28%)</td>
<td>587(54%)</td>
<td>195(18%)</td>
<td>509(87%)</td>
<td>43(7%)</td>
<td>9(2%)</td>
<td>4 (&lt;1%)</td>
<td>4 (&lt;1%)</td>
<td>0</td>
<td>18 (3%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note Percentages are listed in parentheses. Details about each journal articles may be obtained from the first author.

a Journal abbreviations
- ER: Educational Researcher
- JCP: Journal of Counseling Psychology
- JEP: Journal of Educational Psychology
- JRST: Journal of Research in Science Teaching
- JSE: Journal of Special Education
- JSP: Journal of School Psychology
- MLJ: The Modern Language Journal
- RHE: Research in Higher Education
- RME: Journal for Research in Mathematics Education
- TRSE: Theory and Research in Social Education

b Percentages in parenthesis are based on N.
c Percentages in parenthesis are based on n.
d Missing data method abbreviations
- LD: Listwise deletion
- PD: Pairwise deletion
- MS: Mean substitution
- HD: Hot deck
- RE: Regression estimation
- EM: Estimation-maximization
- ML: Maximum-likelihood based
- MI: Multiple imputation
### Table 2A

**Hypothetical Data of Ten Subjects on Three Variables**

<table>
<thead>
<tr>
<th>Subject</th>
<th>( Y ) (Post test)</th>
<th>( X_1 ) (Pre-test)</th>
<th>( X_2 ) (IQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>17</td>
<td>107</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>22</td>
<td>110</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>11</td>
<td>104</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>15</td>
<td>99</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>18</td>
<td>118</td>
</tr>
<tr>
<td>7</td>
<td>26</td>
<td>27</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>25</td>
<td>115</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td>90</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>19</td>
<td>112</td>
</tr>
</tbody>
</table>

### Table 2B

**Monotone Pattern of the Hypothetical Data of Ten Subjects on Three Variables**

<table>
<thead>
<tr>
<th>Subject</th>
<th>( Y ) (Post test)</th>
<th>( X_1 ) (Pre-test)</th>
<th>( X_2 ) (IQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>22</td>
<td>17</td>
<td>107</td>
</tr>
<tr>
<td>2</td>
<td>24</td>
<td>20</td>
<td>100</td>
</tr>
<tr>
<td>4</td>
<td>25</td>
<td>11</td>
<td>104</td>
</tr>
<tr>
<td>5</td>
<td>24</td>
<td>15</td>
<td>99</td>
</tr>
<tr>
<td>7</td>
<td>31</td>
<td>27</td>
<td>122</td>
</tr>
<tr>
<td>8</td>
<td>26</td>
<td>25</td>
<td>115</td>
</tr>
<tr>
<td>10</td>
<td>23</td>
<td>19</td>
<td>112</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>118</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>90</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Comparison of Seven Methods for Handling Missing Data

<table>
<thead>
<tr>
<th>Method</th>
<th>Features</th>
<th>Strength</th>
<th>Weakness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Ad Hoc Methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LD (listwise deletion)</td>
<td>Discards cases with a missing value.  Uses the remaining data to compute results.  A default option in most statistical packages.  Valid only when data are missing completely at random.</td>
<td>Easy.  Does not distort the marginal distribution.</td>
<td>Loss of information  Biased  Inefficient</td>
</tr>
<tr>
<td>PD (pairwise deletion)</td>
<td>Cases with nonmissing values are used to compute means and variances  Pairs of cases with nonmissing values are used to compute correlations and covariances.  Valid only when data are missing completely at random.</td>
<td>Easy.  In AMOS  Efficient  Makes use of all available data  Produces unbiased parameter estimates</td>
<td>Loss of information  Biased  Varied sample sizes  Inefficient</td>
</tr>
<tr>
<td>MS (mean substitution)</td>
<td>Missing values are substituted by means or subgroup means.  Statistical analyses are based on the entire data set.  Valid when data are missing completely at random.</td>
<td>Easy.  In SPSS MVA  Efficient</td>
<td>Biased  Inefficient  Distorts correlations and covariances.  Inefficient</td>
</tr>
<tr>
<td>HD (simple hot-deck)</td>
<td>Missing values are replaced by randomly drawn data already collected in the data set.  Complete data for statistical analyses.  Valid when data are missing completely at random.</td>
<td>Easy.  In SPSS MVA  Efficient</td>
<td>Biased  Inefficient  Distorts correlations and covariances.  Inefficient</td>
</tr>
<tr>
<td>RE (regression estimation)</td>
<td>Missing values are replaced from predicted values in a regression equation.  The regression equation is formed from observations with complete data.  Regression model is specified researchers.  Valid when data are missing at random.</td>
<td>Easy.  In SPSS MVA  Efficient</td>
<td>Model specific or dependent  Failure to specify a reasonable imputation model or to satisfy multivariate normality can bias estimates</td>
</tr>
<tr>
<td><strong>Principled Methods</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIML (full information maximum likelihood)</td>
<td>Available data for each subject used to compute a likelihood and to estimate parameters  Requires specification of an imputation model containing variables believed to be predictive of missingness.  Subject likelihoods are summed and maximum likelihood is used to obtain parameter estimates for the summed likelihood.  Valid when data are multivariate normal and missing data are missing at random.</td>
<td>Efficient  Flexible  Programs available from <a href="http://www.stat.psu.edu/~jls">http://www.stat.psu.edu/~jls</a> or SAS version 9 or later.</td>
<td>Model specific or dependent  Failure to specify a reasonable imputation model or to satisfy multivariate normality can bias estimates</td>
</tr>
<tr>
<td>SPSS’ EM (EM=estimation-maximization)</td>
<td>Each iteration consists of two steps: an E-step followed by an M-step.  Iteratively computes maximum likelihood estimates for parameters.  Iterations continue until the observed log-likelihoods produced in two consecutive iterations are almost identical  Valid when data are missing at random.</td>
<td>Efficient</td>
<td>Time-consuming  Model specific or dependent</td>
</tr>
<tr>
<td>MI (multiple imputation)</td>
<td>Consists of three steps: imputation, analysis and pooling.  Takes into account the uncertainty multiple imputations.  Solves the missing data problem at the beginning of the analysis.  Valid when data are missing at random.</td>
<td>Efficient</td>
<td>Need to specify an imputation model as well as an analysis model.  The third stage—pooling is for parameter estimates and standard errors.</td>
</tr>
</tbody>
</table>

Table 3
Table 4

*Estimates and Standard Errors for Path Coefficients for LD and FIML*

<table>
<thead>
<tr>
<th>Paths</th>
<th>LD</th>
<th>SE</th>
<th>FIML</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>QUALITY → ABILITY</td>
<td>88.87</td>
<td>15.36</td>
<td>155.78</td>
<td>33.71</td>
</tr>
<tr>
<td>QUALITY → COSTS</td>
<td>10.05</td>
<td>8.20</td>
<td>14.99</td>
<td>8.97</td>
</tr>
<tr>
<td>COSTS → PERSIST</td>
<td>-1.44</td>
<td>1.39</td>
<td>-0.57</td>
<td>0.32</td>
</tr>
<tr>
<td>ABILITY → PERSIST</td>
<td>0.06</td>
<td>0.01</td>
<td>0.01</td>
<td>0.009</td>
</tr>
<tr>
<td>QUALITY → PERSIST</td>
<td>3.43</td>
<td>1.09</td>
<td>-3.48</td>
<td>1.23</td>
</tr>
<tr>
<td>QUALITY → %FACULTY Ph.D.</td>
<td>15.94</td>
<td>3.26</td>
<td>18.43</td>
<td>3.49</td>
</tr>
<tr>
<td>ABILITY → AVERAGE TOTAL SAT</td>
<td>1.00</td>
<td>------</td>
<td>1.00</td>
<td>------</td>
</tr>
<tr>
<td>QUALITY → STUDENT to FACULTY RATIO</td>
<td>1.00</td>
<td>------</td>
<td>1.00</td>
<td>------</td>
</tr>
<tr>
<td>QUALITY → AVERAGE EXPENDITURE</td>
<td>13.71</td>
<td>23.68</td>
<td>83.61</td>
<td>33.29</td>
</tr>
<tr>
<td>QUALITY → %FULLTIME</td>
<td>13.97</td>
<td>3.18</td>
<td>22.38</td>
<td>4.27</td>
</tr>
<tr>
<td>QUALITY → No.Admitted /No. Applied</td>
<td>0.06</td>
<td>0.01</td>
<td>-0.14</td>
<td>0.04</td>
</tr>
<tr>
<td>QUALITY → ADDITIONAL FEES</td>
<td>1.00</td>
<td>------</td>
<td>1.00</td>
<td>------</td>
</tr>
<tr>
<td>COSTS → COST BOOKS</td>
<td>2.17</td>
<td>1.55</td>
<td>1.69</td>
<td>0.99</td>
</tr>
<tr>
<td>COSTS → ROOM and BOARD</td>
<td>96.34</td>
<td>42.66</td>
<td>44.52</td>
<td>25.22</td>
</tr>
<tr>
<td>COSTS → OUT of STATE TUITION</td>
<td>38.2</td>
<td>76.12</td>
<td>152.83</td>
<td>86.56</td>
</tr>
<tr>
<td>PERSIST → GRADUATION RATE</td>
<td>1.00</td>
<td>------</td>
<td>1.00</td>
<td>------</td>
</tr>
</tbody>
</table>

Notes

Quality → Ability reflects the direct effect (path) from institutional quality to student ability, LD = listwise deletion, FIML = full information maximum likelihood. Estimates of 1.00 were specified as part of the model.
Table 5

*Logistic Regression Analysis of Adolescent’s Self-inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8.2), n=517*

<table>
<thead>
<tr>
<th>Predictor</th>
<th>β</th>
<th>SEβ</th>
<th>Wald’s χ² (df=1)</th>
<th>p</th>
<th>eβ (odds ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-1.6120</td>
<td>1.3165</td>
<td>1.4991</td>
<td>.2208</td>
<td>Not necessary</td>
</tr>
<tr>
<td>GENDER (boys=1, girls=0)</td>
<td>1.1014</td>
<td>0.2777</td>
<td>15.7291</td>
<td>&lt;.0001</td>
<td>3.0083</td>
</tr>
<tr>
<td>DROPOUT (yes=1, no=0)</td>
<td>2.5369</td>
<td>0.3221</td>
<td>62.0297</td>
<td>&lt;.0001</td>
<td>12.6404</td>
</tr>
<tr>
<td>FAMILY</td>
<td>0.4084</td>
<td>0.1505</td>
<td>7.3601</td>
<td>.0067</td>
<td>1.5044</td>
</tr>
<tr>
<td>EMOTION</td>
<td>0.0081</td>
<td>0.0143</td>
<td>0.3198</td>
<td>.5717</td>
<td>1.0081</td>
</tr>
<tr>
<td>ESTEEM</td>
<td>-0.044</td>
<td>0.0155</td>
<td>8.1678</td>
<td>.0043</td>
<td>0.9570</td>
</tr>
</tbody>
</table>

Overall Model Evaluation

<table>
<thead>
<tr>
<th>Tests</th>
<th>χ²</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio Test</td>
<td>122.04</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score Test</td>
<td>135.07</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald Test</td>
<td>88.03</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Hosmer &amp; Lemeshow Goodness-of-fit Test</td>
<td>5.54</td>
<td>8</td>
<td>.6865</td>
</tr>
</tbody>
</table>

Notes.

Cox and Snell R squared=0.2124. Nagelkerke R squared (Max rescaled R squared)=0.3361. Kendall’s Tau-a = 0.200. Goodman-Kruskal’s Gamma= 0.627.

Somers’ $D_{xy}$ = 0.625. c-statistic = 0.813.
**Logistic Regression Analysis of Adolescent’s Self-inflicting Behavior Risk by SAS® PROC LOGISTIC (version 8.2), n=432**

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$</th>
<th>$SE\beta$</th>
<th>Wald’s $\chi^2$ (df=1)</th>
<th>$p$</th>
<th>$e^\beta$  (odds ratio)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>-1.6341</td>
<td>1.3001</td>
<td>1.5797</td>
<td>.2088</td>
<td>Not necessary</td>
</tr>
<tr>
<td>GENDER (boys=1, girls=0)</td>
<td>1.0470</td>
<td>0.2818</td>
<td>13.8058</td>
<td>.0002</td>
<td>2.8491</td>
</tr>
<tr>
<td>DROPOUT (yes=1, no=0)</td>
<td>2.0772</td>
<td>0.3368</td>
<td>38.0385</td>
<td>&lt;.0001</td>
<td>7.9821</td>
</tr>
<tr>
<td>FAMILY</td>
<td>0.4628</td>
<td>0.1546</td>
<td>8.9597</td>
<td>.0028</td>
<td>1.0648</td>
</tr>
<tr>
<td>EMOTION</td>
<td>0.0065</td>
<td>0.0142</td>
<td>0.2111</td>
<td>.6459</td>
<td>1.0065</td>
</tr>
<tr>
<td>ESTEEM</td>
<td>-0.040</td>
<td>0.0155</td>
<td>6.5224</td>
<td>.0107</td>
<td>0.9608</td>
</tr>
</tbody>
</table>

**Overall Model Evaluation**

<table>
<thead>
<tr>
<th>Tests</th>
<th>$\chi^2$</th>
<th>df</th>
<th>$p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood Ratio Test</td>
<td>80.89</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Score Test</td>
<td>86.24</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Wald Test</td>
<td>63.22</td>
<td>5</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Hosmer &amp; Lemeshow Goodness-of-fit Test</td>
<td>5.23</td>
<td>8</td>
<td>.7322</td>
</tr>
</tbody>
</table>

**Notes.**

Cox and Snell $R$ squared=0.1708. Nagelkerke $R$ squared (Max rescaled $R$ squared)=0.2675. Kendall’s Tau-$a$ = 0.189. Goodman-Kruskal’s Gamma= 0.578. Somers’ $D_{xy}$= 0.576. $c$-statistic = 0.788.
Table 7

*Relative Increase in Variance and Fraction of Missing Information in the Adolescent Data Set*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Relative Increase in Variance</th>
<th>Fraction of Missing Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dropout</td>
<td>0.0270</td>
<td>0.0267</td>
</tr>
<tr>
<td>Behavioral Risk</td>
<td>0.3195</td>
<td>0.2634</td>
</tr>
<tr>
<td>Emotional Risk</td>
<td>0.0435</td>
<td>0.0425</td>
</tr>
</tbody>
</table>
Table 8

Pooled Logistic Regression Analysis Results of Adolescent’s Self-Inflicting Behavioral Risk by SAS® PROC MI and PROC MIANALYZE (version 8.2), n=517

<table>
<thead>
<tr>
<th>Predictor</th>
<th>$\beta$</th>
<th>$SE\beta$</th>
<th>t-test ($df_m)^a$</th>
<th>$p$</th>
<th>95% Confidence Lower Limit</th>
<th>95% Confidence Upper Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONSTANT</td>
<td>−1.9398</td>
<td>1.2052</td>
<td>−1.61</td>
<td>.1075</td>
<td>−4.3021</td>
<td>0.4226</td>
</tr>
<tr>
<td>GENDER (boys=1, girls=0)</td>
<td>1.1860</td>
<td>0.3305</td>
<td>3.59</td>
<td>.0014</td>
<td>0.5068</td>
<td>1.8654</td>
</tr>
<tr>
<td>DROPOUT (yes=1, no=0)</td>
<td>2.0751</td>
<td>0.3215</td>
<td>6.45</td>
<td>&lt;.0001</td>
<td>1.4436</td>
<td>2.7065</td>
</tr>
<tr>
<td>FAMILY</td>
<td>0.4343</td>
<td>0.1450</td>
<td>3.00</td>
<td>.0029</td>
<td>0.1495</td>
<td>0.7192</td>
</tr>
<tr>
<td>EMOTION</td>
<td>0.0090</td>
<td>0.0131</td>
<td>0.69</td>
<td>.4884</td>
<td>−0.0165</td>
<td>0.0346</td>
</tr>
<tr>
<td>ESTEEM</td>
<td>−0.035</td>
<td>0.0142</td>
<td>−2.45</td>
<td>.0142</td>
<td>−0.0627</td>
<td>−0.0070</td>
</tr>
</tbody>
</table>

$^a$ The definition of $df_m$ is given in Equation (12).
Table 9

Percent of Efficiency as a Function of Rate of Missing ($\gamma$) and the Number of Imputations ($m$)

<table>
<thead>
<tr>
<th>$m$</th>
<th>.1</th>
<th>.2</th>
<th>.4</th>
<th>.5</th>
<th>.6</th>
<th>.8</th>
<th>.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>97</td>
<td>94</td>
<td>88</td>
<td>86</td>
<td>83</td>
<td>79</td>
<td>77</td>
</tr>
<tr>
<td>5</td>
<td>98</td>
<td>96</td>
<td>95</td>
<td>91</td>
<td>89</td>
<td>86</td>
<td>85</td>
</tr>
<tr>
<td>10</td>
<td>99</td>
<td>98</td>
<td>96</td>
<td>95</td>
<td>89</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>20</td>
<td>100</td>
<td>99</td>
<td>98</td>
<td>98</td>
<td>97</td>
<td>96</td>
<td>96</td>
</tr>
</tbody>
</table>
Figure 1
Path Model Fitted To Graduation Data