Psychological Methods
2001, Vol. 6, No. 4, 317-329

1082-989X/01/$5.00 DOI: 10.1037//1082-989X.6.4.317

The Use of Multiple Imputation for the Analysis of Missing Data

Sandip Sinharay, Hal S. Stern, and Daniel Russell
Iowa State University

This article provides a comprehensive review of multiple imputation (MI), a technique for analyzing data sets with missing values. Formally, MI is the process of replacing each missing data point with a set of $m > 1$ plausible values to generate m complete data sets. These complete data sets are then analyzed by standard statistical software, and the results combined, to give parameter estimates and standard errors that take into account the uncertainty due to the missing data values. This article introduces the idea behind MI, discusses the advantages of MI over existing techniques for addressing missing data, describes how to do MI for real problems, reviews the software available to implement MI, and discusses the results of a simulation study aimed at finding out how assumptions regarding the imputation model affect the parameter estimates provided by MI.

Researchers in psychology are frequently faced with the problem of incomplete data sets, particularly when there is information about a large number of characteristics for each individual in a study. Sometimes missingness may occur just because of an oversight or data recording error. More often—for example, in a longitudinal study—missing values occur because some individuals are unable to respond one or more times during the study or drop out of the study prematurely. Occasionally, planned missingness is introduced in the design of the study. For example, some measurements may be taken on only a subset of the sample at hand.

Consider a data set used by Lutgendorf et al. (2001) to examine longitudinal predictors of mortality in a sample of 831 individuals over 65 years of age. Baseline measures were taken on this sample in 1981, with assessments of demographic characteristics, physical and mental health status, stress, and social contact. Approximately 6 years later, blood was drawn from these individuals and assays of immunocompetence (based on levels of interleukin-6 [IL-6]) were conducted. Mortality was then monitored over the next 7 years, or 13 years after the baseline interviews.

The focus of the analysis was on whether levels of IL-6 mediated the effects of the psychosocial variables, especially frequency of church attendance, on mortality. One problem with these data involved the relatively high levels of missing information on some of the key predictor variables assessed during the baseline interviews. Of the 831 participants, 266 (32%) had missing data on one or more of these variables. Personal income was the variable with the most missing data, with 113 participants (14%) refusing to indicate their income. Other variables with more than 5% missing data were history of smoking (71 participants, or 9%) and marital status (43 participants, or 5%).

How can someone find the required estimates of the effects of the predictors on the outcome so that the uncertainty due to the missing values is taken into account? Clearly, this is a situation in which missing data could seriously affect the results of the analyses. If we simply eliminate all individuals with missing values (i.e., analyze complete cases only), we will ignore information present in those incomplete observations and may end up with incorrect results. Also, in a multivariate study with a large number of variables, ignoring incomplete observations may result in very few observations being used in the analyses.

Given that this situation is not new to most researchers in psychology, there is a need for a detailed

Sandip Sinharay and Hal S. Stern, Department of Statistics, Iowa State University; Daniel Russell, Department of Psychology, Iowa State University.

Correspondence concerning this article should be addressed to Hal S. Stern, Department of Statistics, 102F Snedecor Hall, Iowa State University, Ames, Iowa 50011-1210. Electronic mail may be sent to hstern@iastate.edu.

and clear discussion of the most recent statistical techniques that can help one to draw inferences on incomplete data sets in an optimal way. The present article is designed to address this need.

Missing Data Mechanisms

There are basically three types of mechanisms that can produce missing values. Missing data are called *missing completely at random* (MCAR) if a missing response occurs purely by chance. Technically, missing data are said to be MCAR if the probability of a missing response is independent of all the (measured and unmeasured) characteristics of the individuals under study.

The second type of missing data mechanism is called *missing at random* (MAR). In this case, missingness does not depend on the missing values. Missingness may, however, depend on other observed characteristics of the individuals. For example, consider a study with income as a key variable of interest. If less educated individuals tend not to report their income, the missing income values may be MAR because whether an individual responds depends on his or her education.

Statistically, the most problematic type of missing data are *missing not at random* (MNAR), or nonignorably missing data, for which missingness is related to the value that would have been observed. In the above example, missing income will be called MNAR if individuals with high or low income tend not to report their income.

When examining a data set, there is no way to distinguish between the MAR and MNAR cases. As we will see later, there are satisfactory techniques for analyzing MAR data with traditional statistical models, but additional modeling is required for analyzing MNAR data.

Recently, there is growing interest in multiple imputation (MI) among researchers in psychology because of the simplicity and generalizability of this method as a tool for analyzing incomplete data sets. Technically, MI is the process of replacing each missing data point with a set of $m > 1$ plausible values to generate m complete data sets. These complete data sets are then analyzed by standard statistical software, and the results are combined to give parameter estimates and standard errors that take into account the uncertainty due to the missing data values.

Existing Approaches to Missing Data

Before discussing MI in detail, we briefly review other methods of analyzing missing data.

Complete Case Analysis

In a complete case analysis of a data set, only individuals with complete information on all the variables under study are included in the statistical analysis. This method, also known as listwise deletion, is very convenient but makes the assumption that individuals with missing information are a random sample from the larger sample of all individuals under study (in which case the missing data are MCAR according to the terminology introduced earlier). If the missing data really are MCAR, this method gives results that are valid (i.e., accurate on average) but inefficient because some information (the incomplete records) is being discarded. In other situations, complete case analysis may lead to inaccurate results. For example, when less educated people are less likely to report their income, complete case analysis (which ignores those people) will result in an average income that is greater than the population parameter.

A related idea is *available case analysis,* which stipulates that individuals with enough information for any calculation are used (e.g., in estimating the correlation between two variables, all individuals with information on that pair of variables are included). This method, also known as pairwise deletion, despite using more information than complete case analysis, is in some sense worse because different estimates are based on different individuals. Also, just as is true of complete case analysis, this method works well when the missing data are MCAR but is inaccurate under any other missing data mechanism.

Maximum Likelihood Estimation and Bayesian Estimation

When confronted with missing values, very accurate results may be obtained by maximum likelihood estimation or Bayesian estimation if one is using a formal probability model (e.g., a normal model) and the missing values are MAR. The probability model is key as both maximum likelihood and Bayesian approaches rely on the complete data likelihood, the function linking the observed and missing data to the model parameters. Inferences are based on the observed data likelihood that links the observed data and the parameters. Formally, this is obtained from the complete data likelihood by adding together the like-

lihood contribution over all possible values of the missing data, also known as integrating out the missing values from the joint density of the observed values and the missing values.

For maximum likelihood estimation, one maximizes the observed data likelihood to obtain the maximum likelihood estimates of the parameters. In Bayesian estimation, the observed data likelihood and a prior distribution for the parameters together give the posterior distribution of the parameters given the observed data. Inferences about the parameters are made by looking at different aspects of the posterior distribution (e.g., the posterior mean of a parameter is one possible estimate of that parameter).

The problem with these methods is that they are model specific; the computations required may differ greatly if one applies different statistical models to the same data set. Also, the computations for these methods may be complicated, especially for nonlinear models. In many problems, the missing values cannot be integrated out analytically and some alternative method has to be used to obtain parameter estimates. It is in this context that the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) may be used for maximum likelihood estimation. The EM algorithm is a method for obtaining maximum likelihood estimates in the presence of missing data. This approach is used in the Amos computer program (Arbuckle, 1997). References to the EM algorithm as a missing data analysis method are not formally correct. In fact, this estimation method is maximum likelihood applied to a specific model.

In Bayesian estimation, computations are again the key issue. Typically, one can use some kind of Markov chain simulation method to generate a sample from the joint distribution of the parameters and the missing data given the observed data (e.g., see Jeon & Stern, 2001). The key idea of Markov chain simulation is to create a random process whose stationary distribution is the distribution of interest and to run the simulation long enough that the distribution of the current draws is close enough to the stationary distribution.

The maximum likelihood and Bayesian methods are useful methods for the analysis of incomplete data provided that the missingness mechanism is MAR. They estimate the parameters of interest without requiring one to fill in the data set. The disadvantage of these methods is that they require fairly sophisticated computational methods and that they are specific to the model being applied. Also, these do not model the missingness mechanism (which is equivalent to assuming that the missing values are MAR) and may give erroneous results if the missingness mechanism is MNAR.

Single Imputation Methods

Single imputation methods are traditional ideas that can be viewed as precursors of MI. These are methods by which each missing value in the data set is replaced by a plausible value and then analyses are carried out using the usual statistical techniques assuming that one has information on all variables for all individuals. In *mean substitution,* all the missing values for a variable are replaced by the mean value of that variable. Although this approach permits the inclusion of all cases in the final analysis, it leads to invalid results. Use of mean substitution will lead to valid estimates of mean values from the data only if the missing values are MCAR, but the estimates of the variance and covariance parameters (and hence correlations, regression coefficients, and other similar parameters) are invalid because this method underestimates the variability among the missing values by replacing them with the corresponding mean. In *regression substitution,* all the missing values of a data set are replaced by the predicted value of that variable from a regression analysis based only on the complete cases. In this method, the mean parameters are correctly estimated for MCAR, but the variance parameters are still underestimated because this method assumes no residual error around the regression line.

A better procedure is *stochastic regression imputation,* in which each missing value in the data set is replaced by a predicted value (as in the regression imputation method) from a regression analysis based on complete cases plus a random residual term. The stochastic regression imputation method will improve on regression imputation if the regression model is reasonable.

In *hot-deck imputation,* each missing value is replaced by an observed value from a randomly chosen case that is similar to the case with the missing value. Similar cases are defined as the set of cases that match the current case on a selected group of covariates. Consider, for example, a situation in which the values of all the characteristics except one are available for Person A. In hot-deck imputation, one finds all individuals in the original data set who have complete information on all the characteristics of interest and who have similar values for the characteristics that were observed for Person A. One then draws an in-

dividual randomly from this set of individuals and replaces the missing value for Person A with the value of the corresponding characteristic of the randomly drawn individual. Although this method preserves marginal distributions, it may distort relationships among the variables. Also, although this is easy to implement for data sets in which only one variable has a missing value, it may be very difficult to implement for the multiple missing variable case because it may be hard to determine from what set of cases to draw the replacement values. The software LISREL 8.3 (Jöreskog, Sörbom, du Toit, & du Toit, 1999) uses a version of the hot-deck imputation method to generate imputations of missing values.

Model-Based MI

History of MI

The use of MI to cope with incomplete data was first proposed by Rubin in 1978 and developed further by Rubin (1987), in the context of large sample surveys in which data collected in a single study are to be used by a potentially large number of investigators for a number of different analyses. However, MI remained obscure and hence unused by nonexperts mainly because of the scarcity of adequate computational facilities. Recently, however, with the advent of faster computers, MI has become quite popular in survey and nonsurvey contexts (Rubin, 1996; Schafer, 1997a; Schafer & Olsen, 1998). Multiple imputation has performed well in a number of recent articles that compare approaches for handling missing data in the structural equation modeling context (Duncan et al., 1998; Gold & Bentler, 2000; Vargas-Chanes, 2000). Another reason for the recent popularity of MI is that after the advent of the EM algorithm in the late 1970s, statisticians began treating missing values as a source of variation to be averaged over (rather than treating missing values as a nuisance). MI can do this averaging in a simple way.

Concept of MI

Basically, MI is an extension of the single imputation idea, whereby each missing value is replaced by a set of $m > 1$ plausible values to generate m apparently complete data sets. These m data sets are then analyzed by standard statistical software, and the results are combined using techniques suggested by Rubin (1987) to give parameter estimates and standard errors that take into account the uncertainty due to the missing data values.

In most applications, just three to five imputations are sufficient to obtain excellent results. Rubin (1987) showed that the efficiency of an estimate based on m imputations is approximately

$$(1 + \gamma/m)^{-1}$$

where γ is the fraction of missing information for the quantity being estimated. For 40% missing information, $m = 5$ imputations give 93% efficiency whereas $m = 10$ imputations increase efficiency only to 96%.

Advantages of MI

As already mentioned, MI was developed in the context of large survey studies in which data collected in a single investigation are to be used by a potentially large number of researchers for a number of different analyses. It is ideal in that context because multiple imputations can be created once by the data collector (who usually has access to more information than individual users) and then all users may analyze the resulting complete data sets using standard statistical software.

This advantage is also true of single imputation methods. However, as noted previously, most single imputation methods have limitations. Also, even if the missing values could be imputed in such a way that the distributions of the variables and the relationships among the variables were not distorted, the imputed data sets obtained by replacing each missing value by some kind of a point estimate would still fail to account for the uncertainty in the missing data and hence would underestimate the variability in the data set. As a result, the standard errors of the parameters would be underestimated and the Type I error rate for any hypothesis test would be higher than the intended rate (i.e., the test would be positively biased). With MI, the results from a series of complete data analyses can be combined to address the uncertainty due to the missing values.

The advantage of MI over maximum likelihood estimation (and Bayesian estimation) is that it is computationally much simpler for most practical situations. The maximum likelihood estimation method is problem specific and may require totally different computational procedures to integrate out the missing data for different models applied to the same data set. By contrast, in MI the same imputed data sets may be used for different types of analyses by different users using any popular statistical software, without any need for these users to worry about addressing the missing data problem.

Assumptions Required by MI

Multiple imputations are generated by assuming a particular imputation model, and the success or failure of MI depends on the propriety of the assumed imputation model. The three aspects of an analysis for which assumptions are required in MI are (a) a model for the data values, (b) a prior distribution for the parameters of the data model, and (c) the nonresponse mechanism.

Data model. The first and the most important step in obtaining multiple imputations for missing values in a data set is to assume a probability model that relates the complete data Y (the combination of the observed values Y_{obs} and the missing values Y_{mis}) to a set of parameters. Using this probability model and the prior distribution on the parameters (to be discussed shortly), one finds a predictive distribution $p(Y_{mis}|Y_{obs})$ for the missing values conditional on the observed values and then generates the imputations from this predictive distribution.

The model assumed should incorporate all the knowledge one has about the process that generated the data. The most convenient model for continuous variables is the multivariate normal assumption. One key advantage is that this model is manageable computationally. It appears (see, e.g., Schafer, 1997a, pp. 147–148, and the references therein) that the multivariate normal model gives quite acceptable results even when the variables are binary or categorical, with the imputations performed assuming a normal model and then the imputed values rounded off to the nearest category. If we have a variable that does not appear normally distributed, it may be transformed to a normal variable and the imputed values transformed back to the original scale. Other models that data analysts have used include a log-linear model for categorical variables, a mixture of a log-linear model and a multivariate normal model for mixed continuous and categorical data sets, and a hierarchical linear model (Bryk & Raudenbush, 1992).

Prior distribution. The statistical approach used to carry out the model-based multiple imputation method is usually Bayesian, and hence we need to specify a prior distribution on the parameters to carry out the analyses. The prior distribution and the complete data model will give us the predictive distribution $p(Y_{mis}|Y_{obs})$ for the missing values conditional on the observed values from which one can generate the imputations. Usually, for convenience, noninformative prior distributions are used to do MI. Because of the subjectivity involved in the choice of the prior distributions, Bayesian methods have at times been criticized. For many data analyses, prior distributions hardly matter because with even moderately large sample sizes any reasonable prior distribution gives essentially the same results. If the sample size is small, then doing the analysis under different prior distributions and seeing whether the results change is a good check before drawing any conclusions.

Missing data mechanism. Model-based MI assumes that the missing values are MAR. The MAR assumption allows one to use the relationships among the variables evident from the observed data to obtain imputed values for the missing observations. For example, in a study in which we have information on many background variables (X) with missing values for a single variable (Y) only, assuming MAR (which basically means that the missingness depends on X only) will lead us to find out how Y depends on X from the complete cases. We then use that dependence relation (most commonly a regression line along with a residual standard error) to predict the missing values of Y from the corresponding X values.

Although it is true that the results of MI may be invalid if the true missing data mechanism is not MAR, this assumption is popular because it is a convenient starting point for data analysis. Addressing the possibility of nonignorably missing data (e.g., by constructing a model to describe the missing data mechanism) will make computations very complicated and will almost certainly be problem specific. It can be a good idea to fit such models as a form of sensitivity analysis, but we do not address that here. Also, if one can collect information on a number of good predictor variables that might govern the missingness mechanism, the MAR assumption (and hence the results from MI) becomes more plausible. This is why the common advice about MI is that one should collect information about any characteristics that might even remotely affect missingness and include those characteristics in the imputation model.

It should be noted that our discussion of missing data mechanisms (e.g., MAR) is in terms of the variables that are included in the data set. If there is a variable which alone governs missingness, we will have an MAR situation if we collect information on that variable. If we have not collected information on that variable, then we may have MCAR, MAR, or MNAR, depending on the relationship of the unmeasured variable to those variables included in the study. Another point with MAR is that it is impossible to check whether this assumption is true. It is not pos-

sible to rule out the possibility that those individuals who did not respond are different in some way without seeing the missing values.

The Imputation Model and the Analysis Model

A key feature of the MI approach is a separation between the model used to obtain the imputations and the final model used for analysis of the data set. Although the imputation is usually done by the person who collected the data, the final analysis may be done by many other users who share the data set. Naturally, the data collector has much better knowledge about the data and is likely the best person to do imputations. Once the data collector has obtained the imputations, future users can use any complete-data technique (which is available in any standard statistical software) to do the final analysis. These users do not need to worry about the missingness mechanism because they basically have access to complete data sets. Rubin (1996) argued that MI is the method of choice when the database constructor and the ultimate users are different entities, with the data constructor having the responsibility to use all of his or her knowledge to obtain suitable imputed values that will give valid results when the ultimate users employ them later.

The imputation model and the data analyses should be compatible to provide good results. Schafer and Olsen (1998) provided an example. If a variable Y is imputed under a normal model that includes the variable X_1 and then the analyst fits a linear regression to predict Y from X_1 and X_2 (where X_2 was not in the imputation model), the estimated coefficient for X_2 would be biased toward zero because the imputation model distorted the relationship between Y and X_2 given X_1.

Conducting Model-Based MI

Mathematical Idea of MI

Rubin (1987) first suggested the model-based MI approach that is the focus of our discussion. Rubin's suggested approach is Bayesian in nature and was made more popular by Schafer (1997a), who provided detailed algorithms for creating multiple imputations in different situations. Our development follows that by Schafer (1997a). The basic idea of MI is to get a predictive distribution for the missing values given the observed data. Let Y be the intended data. Suppose $Y = (Y_{obs}, Y_{mis})$, where the first component denotes the observed part of the data and the second component is the missing data. Assume further that Y follows a distribution $p(Y|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the collection of all the parameters of the model, and also assume that the values are MAR. The predictive distribution, which we denote as $p(Y_{mis}|Y_{obs})$, can be written as

$$p(Y_{mis}|Y_{obs}) = \int p(Y_{mis}, \boldsymbol{\theta}|Y_{obs})d\boldsymbol{\theta}$$
$$= \int p(Y_{mis}|Y_{obs}, \boldsymbol{\theta})p(\boldsymbol{\theta}|Y_{obs})d\boldsymbol{\theta}.$$

We can impute Y_{mis} in two steps by first simulating a parameter value from the observed data posterior $p(\boldsymbol{\theta}|Y_{obs})$ and then simulating a missing data vector from the conditional posterior distribution $p(Y_{mis}|Y_{obs}, \boldsymbol{\theta})$ using the value of $\boldsymbol{\theta}$ generated in the first step. Usually the second step is very easy—for example, generating from a regression model with predictor vector Y_{obs} and parameter $\boldsymbol{\theta}$. The first step, generating a parameter value, requires carrying out a traditional Bayesian analysis with missing data (Gelman, Carlin, Stern, & Rubin, 1995). A variety of approaches are possible, but currently Markov chain simulation methods are often used to do the Bayesian analysis.

In practice, multiple imputations are often generated from the predictive distribution using the first of the two expressions, that is, $p(Y_{mis}|Y_{obs}) = \int p(Y_{mis}, \boldsymbol{\theta}|Y_{obs})d\boldsymbol{\theta}$. The data augmentation (DA) algorithm by Tanner and Wong (1987) is a Markov chain simulation method based on this expression that can be used to simulate from $p(Y_{mis}|Y_{obs})$. At the rth step of this iterative algorithm, one generates missing values from their conditional predictive distribution $Y_{mis}^{(r)} \sim p(Y_{mis}|Y_{obs}, \boldsymbol{\theta}^{(r-1)})$ and then samples the parameter values from a complete-data posterior distribution $\boldsymbol{\theta}^{(r)} \sim p(\boldsymbol{\theta}|Y_{obs}, Y_{mis}^{(r)})$. Starting from an initial value, this forms a Markov chain, $\{(Y_{mis}^{(r)}, \boldsymbol{\theta}^{(r)}); r = 1, 2, \ldots\}$, which converges to the distribution $p(Y_{mis}, \boldsymbol{\theta}|Y_{obs})$. Hence, after convergence of the chain (which occurs after the chain has run for a considerable number of steps), the generated values of Y_{mis} can be viewed as sampled from $p(Y_{mis}|Y_{obs})$. As mentioned before, generating draws of the missing values from $p(Y_{mix}|Y_{obs}, \boldsymbol{\theta}^{(-1)})$ is straightforward. In the second part of the algorithm, generating from $p(\boldsymbol{\theta}|Y_{obs}, Y_{mis}^{(r)})$, the posterior distribution for a complete data set is generally easier to obtain than the posterior distribution $p(\boldsymbol{\theta}|Y_{obs})$ discussed in the previous paragraph.

Imputation With One Missing Variable Under Normality

We provide details for a simple example. Suppose that information is collected on a number of variables

for all individuals in a study. Further suppose that we have missing information on only one variable, say, v. Denote all of the other variables as the vector $\mathbf{u}$. For imputation purposes we assume the joint distribution of all the variables is multivariate normal. Note that we have not yet said anything about the type of analysis that is intended. The joint normality assumption leads one to a normal predictive distribution for v given $\mathbf{u}$. We can write this predictive distribution for the ith individual as $v_i|\mathbf{u}_i \sim N(\mathbf{u}'_i\boldsymbol{\beta}, \sigma^2)$, where the parameter vector $\boldsymbol{\theta}$ of interest for the imputation model is $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$, and $\boldsymbol{\beta}$ has the same number of components as in $\mathbf{u}$. Some prior distribution is required, but as mentioned earlier, should not matter in large samples. Usually, one assumes a noninformative prior distribution for $\boldsymbol{\theta}$, that is, $p(\boldsymbol{\theta}) \propto 1/\sigma^2$. To generate imputations from the predictive distribution, we rely on the basic algorithm described in the previous section. We first need simulations from $p(\boldsymbol{\theta}|Y_{\text{obs}})$. In this simple case, this step is straightforward and does not require any Markov chain simulation method. Let $\mathbf{V}_{\text{obs}}$ denote the vector of observed values of v and $\mathbf{U}$ the fully observed matrix corresponding to the $\mathbf{u}$s.

We follow the Bayesian regression analysis in chapter 8 of Gelman et al. (1995). The complete data posterior distribution of $\boldsymbol{\theta}$ is given by

$$p(\boldsymbol{\theta}|Y_{\text{obs}}) = p(\boldsymbol{\beta}, \sigma^2|\mathbf{V}_{\text{obs}}, \mathbf{U}) = p(\sigma^2|\mathbf{V}_{\text{obs}}, \mathbf{U})p(\boldsymbol{\beta}|\sigma^2, \mathbf{V}_{\text{obs}}, \mathbf{U}).$$

As mentioned, these posterior distributions can be simulated directly without need of the Markov chain simulation method. In fact, the posterior distribution corresponds to the distributional results for a traditional regression analysis based on the complete cases. We can write

$$\mathbf{U} = \begin{pmatrix} \mathbf{U}_{\text{obs}} \\ \mathbf{U}_{\text{mis}} \end{pmatrix} \quad \text{and} \quad \mathbf{V} = \begin{pmatrix} \mathbf{V}_{\text{obs}} \\ \mathbf{V}_{\text{mis}} \end{pmatrix}.$$

The traditional least squares estimates are $\hat{\boldsymbol{\beta}} = (\mathbf{U}'_{\text{obs}}\mathbf{U}_{\text{obs}})^{-1}\mathbf{U}'_{\text{obs}}\mathbf{V}_{\text{obs}}$ and

$$s^2 = \frac{(\mathbf{V}_{\text{obs}} - \mathbf{U}_{\text{obs}}\hat{\boldsymbol{\beta}})'(\mathbf{V}_{\text{obs}} - \mathbf{U}_{\text{obs}}\hat{\boldsymbol{\beta}})}{n_1 - k},$$

where n_1 is the number of complete cases. The components of $p(\boldsymbol{\theta}|Y_{\text{obs}})$ are $p(\sigma^2|\mathbf{V}_{\text{obs}}, \mathbf{U}) = (n_1 - k)s^2/\chi^2_{n_1-k}$ and $p(\boldsymbol{\beta}|\sigma^2, \mathbf{V}_{\text{obs}}, \mathbf{U}) = N(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}, \sigma^2(\mathbf{U}'_{\text{obs}}\mathbf{U}_{\text{obs}})^{-1})$. Given simulations of $\boldsymbol{\beta}$ and σ^2, it is straightforward to generate a simulation of Y_{mis} from the predictive distribution: $p(Y_{\text{mis}}|Y_{\text{obs}}, \boldsymbol{\theta}) \equiv p(\mathbf{V}_{\text{mis}}|\sigma^2, \boldsymbol{\beta}, \mathbf{U}) = N(V_{\text{mis}}|\mathbf{U}_{\text{mis}}\boldsymbol{\beta}, \sigma^2 I)$.

In summary, to create one imputation for the missing responses, the steps are as follows:

1. Generate σ^2 from $p(\sigma^2|\mathbf{V}_{\text{obs}})$.
2. Generate $\boldsymbol{\beta}$ from $p(\boldsymbol{\beta}|\sigma^2, \mathbf{V}_{\text{obs}}, \mathbf{U})$ using σ^2 from the above step.
3. Generate $\mathbf{V}_{\text{mis}}$ from $p(\mathbf{V}_{\text{mis}}|\sigma^2, \boldsymbol{\beta}, \mathbf{U})$ using σ^2 and $\boldsymbol{\beta}$ from the above steps.

Notice that the MI approach assumes a multivariate normal distribution for the variables, which implies a linear regression of v on $\mathbf{u}$ without any interaction terms. So if we have an a priori belief that data analysts will want to consider some interactions among variables, then we need to create and include those interaction terms in our original multivariate normal distribution. Note that though strictly speaking this might violate the multivariate normal assumption, it is plausible that it will still be a reasonable approximation.

We do not provide detailed discussion of imputation when many variables are missing or for nonnormal models. Schafer (1997a, chaps. 6–9) discusses these in detail. Allison (2001) provides an introduction for social scientists. For details about how to proceed with multiple imputation for MNAR situations, interested readers may see Diggle and Kenward (1994), Little (1993), and Little and Rubin (1987, chap. 11).

Combining Estimates

After m imputations have been created for a data set, they may be stored and analyses performed later using any standard statistical package. Because there are now m completed data sets (containing the observed values and the imputed values) instead of one, any statistical procedure has to be done m times, once on each complete data set. The results will differ across the m data sets, reflecting the uncertainty due to the missing observations. To obtain an overall inference, one would have to compute estimates and their standard errors from each of the m completed data sets and then combine them using the rules suggested by Rubin (1987), which are as follows.

Let Q be some function of a parameter of interest (e.g., the population mean or a regression coefficient). Assume Q to be k dimensional. Assume further that, with complete data, inference for Q is based on the statement that $\hat{Q} - Q \sim N(0, D)$, where $\hat{Q}$ is a statistic

estimating Q, and D is a parameter matrix ($k \times k$ dimensional) providing the variance of $\hat{Q}$.

Suppose that m sets of repeated imputations have been drawn and used to form m complete data sets with $\hat{Q}_1, \hat{Q}_2, \ldots, \hat{Q}_m$ as the estimates of Q, and $\hat{D}_1, \hat{D}_2, \ldots, \hat{D}_m$ as the corresponding estimates of D. The average of the m complete-data estimates,

$$\overline{Q} = \frac{\sum_{i=1}^{m} \hat{Q}_i}{m},$$

is the natural estimate for Q. Let

$$\overline{D} = \frac{\sum_{i=1}^{m} \hat{D}_i}{m}$$

be the average of the m complete-data variance estimates, and

$$B = \frac{\sum_{i=1}^{m} (\hat{Q}_i - \overline{Q})^2}{m-1}$$

be the variance between the m complete-data estimates. The quantity

$$T = \overline{D} + \left(1 + \frac{1}{m}\right) B$$

is the total variance of $\overline{Q}$. Inference can be made using $\overline{Q}$, T, and a distributional assumption for some function of them. For example, if Q is a scalar quantity, inference can be made with the assumption that $\overline{Q}$ follows a t distribution with degrees of freedom

$$v = (m-1)\left(1 + \frac{m\overline{D}}{(m+1)B}\right)^2;$$

for example, $\overline{Q} \pm t_{\alpha;c} \sqrt{T}$ will provide an approximate $100(1 - \alpha)\%$ confidence interval for Q.

Software for Performing MI

Schafer (1997b, 1997c, 1997d, 1997e, 1999) has written general-purpose MI software for the analysis of missing data. NORM (Schafer, 1999) is a standalone application for PCs running Windows 95, 98, and NT that performs MI under a multivariate normal model. The program may be downloaded free of charge from his Web site (see the Reference List). Also available at the same Web site are four different packages for creating multiple imputations in S-PLUS (Insightful Corporation, 2001): NORM (Schafer, 1997d), which performs MI under a multivariate normal model; CAT (Schafer, 1997b), for multivariate categorical data under log-linear models; MIX (Schafer, 1997c), for mixed data sets containing both continuous and categorical data under the general location model; and PAN (Schafer, 1997e), for multivariate panel data or clustered data under a multivariate linear mixed-effects model. All four packages are available as libraries of functions to be run in S-PLUS (Insightful Corporation, 2001). For efficiency, the computationally intensive portions are performed in Fortran-77; the compiled Fortran object code is dynamically loaded in the S-PLUS session. Recently, some of Schafer's programs have been incorporated into the S-PLUS package sold commercially (Insightful Corporation, 2001).

The multiple imputation procedure MI in SAS Version 8.02 also creates multiply imputed data sets for incomplete multivariate data. Once the m complete data sets are analyzed using a standard SAS procedure, the MIANALYZE procedure can be used to combine the results, using the method of Rubin (1987). These two procedures are available in experimental form in Release 8.1 of the SAS system. In procedure MI, the default option for the method to impute the missing values is the multivariate normal model (Rubin, 1987) algorithm. Other options are available as well.

The Windows program SOLAS performs MI as well. Version 1.1 of SOLAS (Statistical Solutions, Inc., 1998), which used propensity scores and an "approximate Bayesian bootstrap" (Rubin, 1987; Rubin & Schenker, 1986), was found to give very unsatisfactory results (see, e.g., Allison, 2000). However, SOLAS 3.0 (Statistical Solutions, Inc., 2001) includes model-based MI (Rubin, 1987) along with the older method.

An Example of an MI Analysis

We now discuss in detail how to do MI for an actual data set using standard software. To do this we use the data set collected by Lutgendorf et al. (2001) discussed earlier.

Description

The data set contains 28 variables for 831 individuals with some values missing for some individuals. We omit the variable reflecting mortality and work with the remaining 27 measures. Table 1 shows a part of the data set with missing values marked by "?." As

Table 1
Part of the Lutgendorf et al. (2001) Data Set

IL-6 scores	Age (years)	Education (years)	Gender	Marital status	Income	...	Church attendance
0.90	82	16	1	1	5	...	2
0.25	82	10	1	0	?	...	1
0.00	66	16	0	1	6	...	5
−0.14	69	12	1	1	2	...	3
?	80	13	1	1	?	...	4
0.64	76	10	0	1	5	...	5
0.21	65	?	0	1	?	...	1
0.45	66	12	1	1	4	...	1
0.48	65	7	0	?	2	...	5
?	68	12	1	1	4	...	6

Note. Variables were coded as follows: Gender: 1 = female; 0 = male. Marital status: 1 = married; 0 = not married. Income: 1 = less than $2,000; 2 = $2,000–$4,999; 3 = $5,000–$6,999; 4 = $7,000–$9,999; 5 = $10,000–$14,999; and 6 = over $15,000 (in 1981 U.S. dollars). Church attendance reflects frequency of attending religious activities: 1 = never; 2 = 1 to 2 times per year; 3 = every few months; 4 = 1 to 2 times per month; 5 = once per week; and 6 = more than once per week. IL-6 = interleukin-6 levels. ? = missing data.

we can see, there are missing values for more than one variable in the data set.

The statistical analysis of interest is the regression analysis of the IL-6 scores on the 26 predictor variables. However, because of the presence of missing values, the question is how to compute the regression coefficients and assess their significance taking into account the uncertainty involved with the missing observations.

The first step in conducting an MI analysis involves deciding about the imputation model, which means deciding which variables to include in performing the imputation. One should pay attention to our previous discussion in *The Imputation Model and the Analysis Model* section before selecting a model. In this case, we opted to include all 27 variables in the imputation model to make it compatible with the analysis model (in which IL-6 scores are regressed on the other 26 variables). Furthermore, it was assumed that the variables are jointly normally distributed, and hence we used Schafer's NORM package. The NORM package is easy to use, requiring the data analyst to complete a number of dialogue boxes. Our use of the package is described here. Additional details can be found at the Web site www.public.iastate.edu/~hstern.

Step 2 involved preparing the data for input into the NORM program. The program will read data in an ASCII format, with blank spaces between data elements. Importantly, a single value needs to be assigned to all missing data; leaving a blank when a value is missing will not work given that the program is reading the data in free format. Any unique value for the missing data will work; the program asks for the value used (and assumes −9 as the default value).

Step 3 involves reading the data into the NORM program. To do this, one opens a session and indicates the name of the data file and the missing value code. One can also provide names for the variables, which helps in reading the output. The program will provide summary statistics for the variables, such as means and standard deviations as well as the number and percentage of cases that are missing information for each variable. Plots of the distribution of scores on each variable can also be generated, which helps in identifying measures that are not normally distributed. Also available from within the program are various transformation procedures for addressing nonnormality. An important issue to consider before proceeding with the MI analysis concerns whether to round the imputed values that are derived from the program. For example, consider the variable income from the current example, which had a large proportion of missing data. Values on this variable ranged from 1 to 6, reflecting different income groups (see Table 2). Imputed values from the multivariate normal model will not, however, conform to this range of values. Some individuals with missing data on income may receive imputed values that are out of range (e.g., <1 or >6). The NORM program includes various options for dealing with this issue. In addition to not doing any rounding at all, one can also round to one of the observed values, which addresses the problem of both out of range values and noninteger values for integer variables. The user chooses how to address this issue

Table 2
Imputed Income Values

Case	Imputation 1	Imputation 2	Imputation 3	Imputation 4	Imputation 5
1	5	6	2	1	4
2	5	3	6	3	4
3	5	2	6	4	2
4	4	6	4	1	5
5	6	5	6	4	6
6	6	6	6	3	6
7	3	3	3	3	2
8	2	5	4	5	5
9	6	6	6	6	4
10	6	3	4	6	6

Note. Income was coded so that 1 = less than \$2,000; 2 = \$2,000–\$4,999; 3 = \$5,000–\$6,999; 4 = \$7,000–\$9,999; 5 = \$10,000–\$14,999; and 6 = over \$15,000 (in 1981 U.S. dollars).

in the imputation for each variable included in the analysis. For example, we chose to use the observed value option for income (so that imputed values are rounded to the nearest observed values), whereas no rounding at all was done for values of IL-6 that were missing.

Step 4 in the MI analysis involves finding starting values for the data augmentation algorithm used to generate the imputations. NORM takes the starting value to be the maximum likelihood estimates (under a joint normal model) of the mean vector and the covariance matrix of the variables in the imputation model. The maximum likelihood estimates are obtained using an EM algorithm. In this case, the EM algorithm converged after 50 iterations, with the output saved in a file for review and printing.

Step 5 involves the data augmentation procedure, in which the imputed data sets are actually generated by the program. Schafer (1997a) recommended that 5 to 10 imputed data sets be generated and used in subsequent analyses. Furthermore, proper multiple imputations are independent draws of missing data from their predictive distribution. Hence, to ensure that the multiple imputations are truly independent, it is recommended that each imputed data set be generated after every X iterations, with a value of X chosen that is at least twice the number of iterations that led to convergence under the EM algorithm. Therefore, we set the program to conduct 500 iterations, with an imputed data set generated after each 100 iterations.

To illustrate the results of these imputations, Table 2 presents the imputed values for income for the first 10 cases that had missing data on that variable from these five imputed data sets. It is evident that for some individual participants, the imputed income values vary greatly from imputation to imputation, whereas for other individuals there is little variability in the imputed values of income.

The five complete data sets (incorporating observed and imputed values) are then read into the statistical package that would be used in conducting the regression analyses. It should be noted that in these files, all 831 participants have data on all of the variables as a result of the imputation process. Once these data sets have been created, we conduct the analyses regressing IL-6 scores on the 26 predictor variables. This results in five sets of 27 regression coefficients (for the intercept term and the 26 predictor variables) and 27 standard errors. The NORM program also includes an option for reading in these results of the five sets of analyses and combining them together to generate average parameter estimates (regression coefficients in this case) and pooled estimates of the standard errors using the procedures described above. We therefore created an ASCII file containing these five sets of regression coefficients and standard errors and read this file along with a file of names for the predictor variables included in the analysis.

Table 3 presents the pooled estimates based on the five imputed data sets (N = 831). The table also includes the estimates obtained from a regression analysis using the cases (565 out of 831) with complete data only. The effects that are statistically significant at the 5% level are marked with an asterisk. As we can see, the results are very similar across the two sets of analyses, though social support is no longer a significant predictor following imputation. It should also be noted that the proportion of explained variance (R^2) in IL-6 does not vary greatly across the two sets of analyses. For the complete cases analysis, R^2 = .11, whereas for the analyses based on the imputed data, the range of R^2 values is from .10 to .11.

Table 3
Estimates Obtained by Different Techniques for Analyzing Missing Data

Predictor variable	Complete cases		Multiple imputation	
	Coefficient	*SE*	Coefficient	*SE*
Intercept	−.283	.463	−.038	.290
Age	.008*	.003	.007*	.002
Education	.001	.005	−.001	.004
Female	−.017	.031	−.042	.027
Married	.004	.030	.018	.025
Income	.001	.010	−.007	.009
Mental status	−.021	.014	−.012	.011
Morale	−.001	.002	.002	.002
Health status	.005	.003	.003	.003
BMI	.012*	.003	.009*	.002
Stress	.004	.008	.002	.006
Loneliness	−.010	.012	−.010	.009
Club involvement	.002	.005	−.002	.004
Network size	.000	.001	−.001	.001
Social support	−.012*	.006	−.001	.005
Church attendance	−.025*	.009	−.019*	.007

Note. Coefficients are standardized. Also included in the regression analysis were history of diabetes, heart disease, and cancer, as well as measures of smoking, alcohol consumption, and problems sleeping. None of these control variables were statistically significant. BMI = body mass index.
* $p < .05$.

One key benefit of imputation is that the standard errors of the coefficients are smaller than in the complete case analysis. The standard errors are not quite as small as they would be if the original data set were completely observed, but MI retrieves some of the lost information.

Effects of the Imputation Model on Parameter Estimates

To examine how the assumptions of the imputation model affect the estimates obtained by MI, we performed a simulation study. The details can be found in Russell, Stern, and Sinharay's (2001) study; we provide the important findings of the study here.

The simulations varied the missing data mechanism (MAR, MNAR), the specific percentage of missing data (from 10% to 70%), the true value of the parameter of interest (the mean income and the correlation between income and education), and the imputation model (multivariate normal with varying sets of covariates). For each simulation scenario, 10,000 data sets were constructed and analyzed as follows:

1. We generated a data set having the same variables as the IL-6 data set using a joint multivariate normal model with the same mean and variance as our original data set, except that we varied the correlation between income and education.
2. We identified a subset of the values of the variable income as missing and ignored them. The choice of the missing cases was random, depending on the missing data mechanism used and the percentage of missing cases desired.
3. We used model-based MI to create five imputed data sets. We then computed and averaged the values of the statistics of interest (corresponding to parameters in the model) over these imputed data sets.

The mean and variance of the 10,000 averages were computed. The former gives us a point estimate of the parameter of interest (in this case, the population value that is known to us), whereas the standard error (which is the square root of the variance divided by 100, the square root of 10,000) gives us an idea about the precision of the imputation estimate of the parameter of interest. An estimate was judged to be biased if it was more than two standard errors away from the true parameter value.

Our parameters of interest are the mean of income and the correlation between income and education (this is the highest correlation in the data set involving income). We conducted simulations under the missing data mechanisms MAR and MNAR and varied the proportion of missing incomes (10%, 40%, and 70%) and the number of covariates used to impute income (2, 3, 10, and 20, where the set of 2 is nested within the set of 3, and so on). We fixed the correlation coefficient between income and education at 0, 0.3, 0.6, and 0.8. In addition to performing MI, we also estimated the parameters of interest using complete case analysis.

For the mean of income under the MAR mechanism, the estimates obtained by complete case analysis were biased (with the extent of bias increasing as the proportion missing increased), whereas estimates obtained by MI were unbiased. For the mean parameter and MNAR, the estimates obtained by both methods were biased, although those obtained by MI were more accurate. Also, the higher the multiple correlation between income and the other covariates (because of either higher correlation between income and education or because of the use of a larger number of covariates), the less was the extent of bias for MI. As the multiple correlation became very high, the covariates predicted income very well; hence, missingness

can be viewed as dependent on the covariates, which means we have data that are approximately MAR.

For the correlation coefficient and MAR, the estimates obtained by complete case analysis were biased (as was true for the mean). There is no bias in the estimated correlation coefficient when MI is used with only one covariate, but the estimates are biased negatively in almost all cases when 20 covariates are used. Looking deeper into the problem, we determined that the variance of income was being overestimated for these situations, causing an underestimation of the correlation coefficient. This suggests that there is, in fact, a tension regarding the number of covariates included in the imputation model. Having more covariates improves the chance that MAR will be plausible but also introduces more parameters and thus more variability.

For the correlation coefficient and MNAR, we saw the same patterns as in the case with the mean parameter and MNAR—the MI estimates were less biased than the complete case analysis estimates, with the extent of bias diminishing with increasing multiple correlations. Also, we found that the reduction of bias with inclusion of more covariates was larger than the increase in bias we saw in the MAR case. Because one can never be sure whether the situation is MAR or MNAR, these simulation results suggest erring on the side of including too many covariates in the imputation model rather than too few.

Our simulations also show, as expected, that the effectiveness of MI depends on the imputation model assumed (e.g., on the number of covariates used in the model) and the strength of the relationship between the variables. They show that if we can model the data correctly and if we have strong associations between the variables, we can obtain good results using MI irrespective of the missing data mechanism. The latter is a very big advantage of MI.

Discussion

MI addresses the missingness issue in the proper way—it takes into account the uncertainty in the missing values. Hence, it is an improvement over most of the existing methods for dealing with missing data. If a standard maximum likelihood or Bayesian analysis under a model of interest can be done easily for a data set with missing values, then it is preferable to MI because the former is more efficient. However, maximum likelihood estimation or Bayesian analysis can be difficult for many problems (e.g., generalized linear models) whereas MI is a flexible and general approach.

Model assumptions made in imputing the missing values are important. The recommendation here is that the data collector should use all of his or her knowledge (and that of experts in the field) about the problem to impute the missing values because any discrepancy between the imputation model and the analysis model may give rise to unreliable estimates. In forming the imputation model, one should include as much reasonable covariate information as is available. We saw from our simulation that even if the missingness mechanism is MNAR, MI may give quite reasonable estimates if there are strong covariates.

In spite of all its advantages, one has to remember that MI is not a miracle cure for all difficulties. For example, standard software for creating multiple imputations always assumes that the missingness is MAR. Until now, there have been no principled MI methods for the MNAR situation readily available to data analysts. If the true missingness mechanism is MNAR, MI software will provide improved but still biased estimates. This is a potential problem for all researchers, for in an actual situation it is never known whether the missingness mechanism is MAR or MNAR. So our principal recommendation is to use MI with the MAR assumption while being aware of its limitations. Even if the MAR assumption is incorrect, MI with a set of good covariates may produce estimates with little bias and will be more convenient than a complicated MNAR procedure. When possible a sensitivity analysis that considers plausible MNAR models could supplement an MI analysis under the MAR assumption.

References

Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28,* 301–309.

Allison, P. D. (2001). *Missing data.* Thousand Oaks, CA: Sage.

Arbuckle, J. L. (1997). *Amos users' guide (Version 3.6).* Chicago: SmallWaters Corporation.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

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

Diggle, P. J., & Kenward, M. G. (1994). Informative drop-

out in longitudinal data analysis. *Applied Statistics, 43,* 49–73.

Duncan, T. E., Duncan, S. C., & Li, F. (1998). A comparison of model- and multiple imputation-based approaches to longitudinal analyses with partial missingness. *Structural Equation Modeling, 5,* 1–21.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (1995). *Bayesian data analysis.* London: Chapman & Hall.

Gold, M. S., & Bentler, P. M. (2000). Treatments of missing data: A Monte Carlo comparison of RBHDI, interactive stochastic regression imputation and expectation-maximization. *Structural Equation Modeling, 7,* 319–355.

Insightful Corporation. (2001). S-PLUS 6 for Windows [Computer software]. Seattle, WA: Author.

Jeon, Y., & Stern, H. S. (2001). *Bayesian inference for structural equation models with incomplete data.* Manuscript submitted for publication.

Jöreskog, K., Sörbom, D., du Toit, S., & du Toit, M. (1999). *LISREL 8: New statistical features.* Chicago: Scientific Software International.

Little, R. J. A. (1993). Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association, 88,* 125–134.

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

Lutgendorf, S. K., Russell, D. W., Ullrich, P., Harris, T., de la Mora, A., & Wallace, R. (2001). *IL-6 mediates beneficial effects of religious attendance on mortality in older adults.* Unpublished manuscript.

Rubin, D. B. (1978). Multiple imputations in sample surveys—A phenomenological Bayesian approach to nonresponse. In *Proceedings of the Survey Research Methods Section* (pp. 20–34). Alexandria, VA: American Statistical Association.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* New York: Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association, 90,* 822–828.

Rubin, D. B., & Schenker, N. (1986). Multiple imputation for interval estimation from simple random samples with ignorable nonresponse. *Journal of the American Statistical Association, 81,* 366–374.

Russell, D. W., Stern, H., & Sinharay, S. (2001). *An evaluation of multiple imputation as an approach to missing data.* Manuscript submitted for publication.

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

Schafer, J. L. (1997b). CAT Library for S-PLUS [Computer software]. Retrieved July 18, 2001, from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L. (1997c). MIX Library for S-PLUS [Computer software]. Retrieved July 18, 2001, from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L. (1997d). NORM Library for S-PLUS [Computer software]. Retrieved July 18, 2001, from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L. (1997e). PAN Library for S-PLUS [Computer software]. Retrieved July 18, 2001, from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L. (1999). NORM (Version 2.03 for Windows) [Computer software]. Retrieved July 18, 2001, from http://www.stat.psu.edu/~jls/misoftwa.html

Schafer, J. L., & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research, 33,* 545–571.

Statistical Solutions, Inc. (1998). SOLAS Version 1.1 [Computer software]. Cork, Ireland: Author.

Statistical Solutions, Inc. (2001). SOLAS Version 3.0 [Computer software]. Cork, Ireland: Author.

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

Vargas-Chanes, D. (2000). *Imputation methods for incomplete panel data with applications to latent growth curves.* Unpublished doctoral dissertation, Iowa State University, Ames, Iowa.

Received April 4, 2000
Revision received August 15, 2001
Accepted August 17, 2001 ■