Review of Educational Research
Winter 2004, Vol. 74, No. 4, pp. 525–556

Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement

James L. Peugh
University of Nebraska
Craig K. Enders
University of Nebraska

Missing data analyses have received considerable recent attention in the methodological literature, and two "modern" methods, multiple imputation and maximum likelihood estimation, are recommended. The goals of this article are to (a) provide an overview of missing-data theory, maximum likelihood estimation, and multiple imputation; (b) conduct a methodological review of missing-data reporting practices in 23 applied research journals; and (c) provide a demonstration of multiple imputation and maximum likelihood estimation using the Longitudinal Study of American Youth data. The results indicated that explicit discussions of missing data increased substantially between 1999 and 2003, but the use of maximum likelihood estimation or multiple imputation was rare; the studies relied almost exclusively on listwise and pairwise deletion.

KEYWORDS: EM algorithm, maximum likelihood estimation, missing data, multiple imputation, NORM.

Missing data are a common problem in quantitative research studies. Standard statistical procedures were developed for complete data sets, so missing values represent a considerable nuisance to the analyst. Traditionally, missing data were dealt with by means of various ad hoc methods that attempted to "fix" the data before analysis. The blanket removal of cases with missing data (i.e., listwise deletion) is one such strategy. Another method involves substituting missing values with the variable mean. Unfortunately, these ad hoc traditional methods can seriously bias sample statistics and have been criticized in the methodological literature. Referring to those methods, Little and Rubin (1987) stated, "[W]e do not generally recommend any of them" (p. 39). More recently, a report by the APA (American Psychological Association) Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999) strongly discouraged the use traditional missing-data methods (e.g., listwise and pairwise deletion), stating that they are "among the worst methods available for practical applications" (p. 598).

During the last 25 years, there have been substantial methodological advances in the area of missing-data analyses (West, 2001). Specifically, two so-called

"modern" missing-data techniques, maximum likelihood (ML) estimation and multiple imputation (MI), are currently considered "state of the art" (Schafer & Graham, 2002) and are the recommended procedures in the methodological literature. These methods have a strong theoretical framework and are supported by a growing number of empirical research studies that demonstrate their effectiveness (e.g., Arbuckle, 1996; Enders, 2001a, 2001b, 2003; Enders & Bandalos, 2001; Gold & Bentler, 2000; Graham, Hofer, & MacKinnon, 1996; Graham & Schafer, 1999; Muthén, Kaplan, & Hollis, 1987; Kaplan, 1995; Wothke, 2000). Furthermore, the availability of ML and MI routines has increased dramatically in the last few years, and these methods are now implemented in a variety of commercial and freeware software packages.

Quoted in a recent issue of APA's *Monitor on Psychology,* Stephen West, the editor of *Psychological Methods,* stated that "[r]outine implementation of these new methods of addressing missing data [ML and MI] will be one of the major changes in research over the next decade" (Azar, 2002, p. 70). In line with that prediction, the purpose of this article is threefold. First, we provide an overview of Rubin's (1976) theoretical framework for missing data and show how it supports the use of ML and MI. Although these procedures may be familiar to many methodologists, our informal observations (e.g., discussions with colleagues and manuscript reviews) underscore the need to disseminate the information in a nontechnical fashion to substantive researchers. Second, if missing-data estimation will undergo a major change during the course of the coming decade, one might reasonably ask, How do researchers currently deal with missing data in their studies? Accordingly, the second purpose of this article is to provide a methodological review of current missing-data reporting practices in a sample of educational and applied psychology journals from 1999 and 2003. Finally, we conducted a heuristic analysis to demonstrate the use of ML and MI, in hopes of providing a procedural model for substantive researchers as they begin to implement these techniques in their own work.

An Overview of Missing-Data Issues

Missing-Data Theory

To appreciate why ML and MI are preferred over traditional missing-data methods, it is first necessary to understand Rubin's (1976) theoretical framework for missing data. Rubin outlined three "mechanisms" that can be thought of as probabilistic explanations for why data are missing. In another sense, these mechanisms represent assumptions that dictate the conditions under which a particular missing-data method will provide optimal performance. Although the subsequent discussion may give the impression that Rubin's mechanisms are mutually exclusive, note that all three may be present in a given data set (Yuan & Bentler, 2000).

Rubin (1976) defined a *missing completely at random* (MCAR) mechanism as one where the missing values on a particular variable X are unrelated to other variables in the data set as well as the underlying values of X itself. Essentially, the observed data points represent a random sample of the hypothetically complete data set. To illustrate, suppose an educational researcher is conducting a longitudinal study of reading achievement in an elementary school population. Data would be described as MCAR if children were absent from an assessment because of random factors such as illness or a death in the family. Similarly, a

child might permanently leave the study if her parents relocated to a different city. Assuming these factors were unrelated to other measured variables such as socioeconomic status, the observed scores represent a random sample of the hypothetically complete data set.

In certain circumstances MCAR missing data might even be a purposive byproduct of the data collection procedure (Graham, Hofer, & MacKinnon, 1996; Graham, Taylor, & Cumsille, 2001; Kaplan, 1995). For example, the National Assessment of Educational Progress (NAEP) uses a matrix sampling procedure to administer different blocks of items to examinees. By design, examinees will have complete data on the blocks of items that were administered, and missing values on the blocks that were not administered. The resulting data are MCAR, as the missing item blocks are, by definition, unrelated to a student's underlying achievement as well as other measured variables. It is important to note that MCAR is the only mechanism that can be empirically tested from a set of data (Little, 1988).

Rubin's *missing at random* (MAR) mechanism is less restrictive in the sense that missing values on a variable X can be related to other measured variables but still must be unrelated to the underlying values of X. Continuing with the reading assessment example, suppose that children from low-income households (e.g., students who receive free or reduced-price lunch) are found to have higher rates of attrition than other students. Furthermore, within a given income bracket, there is no relationship between attrition and achievement (i.e., there is no residual relationship between attrition and achievement once income is controlled for). In this case the propensity for missing data is related to a measured variable (income) but unrelated to a student's underlying achievement level. Note that MAR is an untestable assumption and could only be verified if we had knowledge of the missing achievement scores.

Finally, a *missing not at random* (MNAR) mechanism results when the probability of missing values on a variable X is related to the underlying values of X. Returning to the reading assessment example, the data would be described as MNAR if children who possessed poor reading skills were more likely to skip test questions because of comprehension difficulties. In this case, missing values on the reading assessment are directly related to underlying reading achievement.

Rubin's (1976) missing-data mechanisms have important implications for the performance of a given missing-data method. With any statistical procedure, the quality of the inferences we make is, in part, a function of meeting certain analytic assumptions (e.g., homogeneity of variance in ANOVA analyses). ML and MI require the MAR assumption and thus will produce parameter estimates (e.g., regression weights) that are unbiased and efficient (i.e., have low sampling fluctuation) when data are MAR or MCAR. In contrast, traditional missing-data methods (e.g., listwise deletion) will generally produce biased parameter estimates under MAR and generally require MCAR data.

When choosing a statistical procedure, we often prefer methods that are robust to, or minimally affected by, assumption violations. If Rubin's (1976) mechanisms are viewed as assumptions (albeit largely untestable), a compelling argument can be made that ML and MI are more "robust" in the sense that they perform optimally under a wider variety of conditions (MAR and MCAR) than traditional methods (typically, MCAR only). However, as in many other facets of life, there is no such thing as a (statistical) free lunch. The previous discussion is not meant

to imply that ML and MI will always provide optimal performance, as those methods will be biased if data are MNAR. However, some authors have suggested that the bias may be less than that associated with traditional approaches in many cases (e.g., Muthén, Kaplan, & Hollis, 1987). Second, ML and MI will always require the multivariate normality assumption, even in cases where an alternative analytic procedure may not. However, there is evidence to suggest that normality violations minimally affect parameter estimate bias (Enders, 2001a; Graham & Schafer, 1999). Based on the methodological literature to date, we believe that, in many cases, violating missing-data assumptions (e.g., using listwise deletion when data are MAR) will be far more deleterious than violating the multivariate normality assumption required by ML and MI.

Traditional Missing-Data Techniques

Literally dozens of ad hoc missing-data techniques have been proposed in the literature. In the interest of space, we limit our discussion to a small selection of these techniques. A number of excellent sources are available to readers who are interested in learning more about traditional missing-data methods (e.g., Allison, 2002; Little & Rubin, 2002; Schafer & Graham, 2002).

Listwise Deletion

Listwise deletion discards all cases with missing values on one or more variables. Not surprisingly, this can result in a potentially dramatic reduction in the sample size, and thus in statistical power. Perhaps more problematic is the fact that, in general, listwise deletion will produce unbiased parameter estimates only when data are MCAR. Even when data are MCAR, the reduction in sample size results in lower power than would be obtained from ML and MI (e.g., Enders & Bandalos, 2001).

Pairwise Deletion

Pairwise deletion attempts to use all available data by discarding cases on an analysis-by-analysis basis. This method is frequently described in the context of a covariance (or correlation) matrix, whereby each variance and covariance term is computed by using all cases with complete data on a given variable or variable pair. However, the definition of pairwise deletion need not be restricted to correlational analyses. Our methodological review revealed numerous situations where a series of ANOVA analyses were conducted, each based on a different n.

Pairwise deletion also has important limitations. The comparability of analyses within a study is problematic, as different subsets of cases are used for each analysis. It is also widely documented that a pairwise deletion covariance matrix need not be positive definite (i.e., certain elements in the matrix may take on impossible values, given the other elements), which can cause estimation problems for multivariate analyses that rely on a covariance matrix (e.g., structural equation models). Finally, pairwise deletion also requires the MCAR assumption to produce unbiased parameter estimates.

Mean Imputation

Several variations of mean imputation (i.e., mean substitution) have been proposed. Typically, the arithmetic mean of each variable is computed from the available scores, and missing values are replaced by the means. The simplicity of this method

is appealing, as the "filled-in" data are analyzed as if there were no missing values to begin with. However, imputing missing values with the mean will truncate the variance of the variable as well as its covariance with other variables. This is clear if one examines the following formula for a covariance:

$$\text{cov}_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{n - 1} \quad (1)$$

It is straightforward to see that any case with a missing score (X or Y) will contribute a value of zero to the numerator of Equation 1, thereby reducing the magnitude of the covariance, and by extension, the Pearson's correlation ($r = \text{cov}_{xy}/\sigma_x\sigma_y$). The missing-data mechanism matters very little in this case, as mean imputation will produce biased estimates of any parameter except the mean, regardless of whether data are MCAR, MAR, or MNAR.

Regression Imputation

Also referred to as conditional mean imputation, regression imputation replaces missing values with the predicted scores from a linear regression equation. Regression imputation is relatively straightforward if missing values are isolated on a single variable (i.e., there is a single, univariate missing-data pattern). In this case the incomplete variable is regressed on other measured variables, and missing values are replaced with the predicted scores from this analysis. The problem with this procedure is that the "filled-in" data lack variability present in the hypothetically complete data set, because all imputed values fall directly on a regression line. The resulting bias in variance and covariance terms can be mitigated by adding a randomly sampled residual term to each imputed value (i.e., stochastic regression imputation). Regression imputation can become fairly complicated when there are multiple patterns of missing data, as different regression equations must be constructed for each unique pattern. Regression imputation can produce parameter estimates that are consistent (i.e., approximately unbiased in large samples) under MAR; Little and Rubin (2002, p. 64) describe this method as a "historical precursor" to ML estimation described subsequently.

Modern Missing-Data Techniques

We now describe the so-called "modern" missing-data techniques currently recommended in the methodological literature, ML and MI.

Maximum Likelihood Estimation

Many widely used statistical procedures (e.g., structural equation models and hierarchical linear models) rely on ML estimation, rather than least squares, to obtain estimates of model parameters. The basic goal of ML estimation is to identify the population parameter values most likely to have produced a particular sample of data. This usually requires an iterative process whereby the model fitting program "tries out" different values for the parameters of interest (e.g., regression coefficients) en route to identifying the values most likely to have produced the sample data. The fit of the data to a particular set of parameter values is gauged by a log likelihood value that quantifies the relative probability of a particular sample,

given that the data originated from a normally distributed population. Interested readers can consult Enders (in press) for an overview of the basic principles of ML estimation. To be clear, ML estimation can be used to estimate models with or without missing, and could be used in conjunction with a traditional missing-data technique (e.g., a structural equation model could be estimated following listwise deletion). However, throughout this article we use the term ML to refer to maximum likelihood estimation with missing data under the MAR assumption (also referred to as *direct ML* or *full information ML* in the missing-data literature).

To illustrate ML estimation, suppose it was of interest to estimate a vector of means and a covariance matrix ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively) from an incomplete data set. ML estimation involves the computation of a log likelihood value at each iteration, or estimation cycle, and an individual's contribution to this log likelihood is shown in Equation 2.

$$\log L_i = K_i - \frac{1}{2}\log\left|\boldsymbol{\Sigma}_i\right| \frac{1}{2}(\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) \tag{2}$$

The scores for individual i are substituted into $\mathbf{y}_i$, and the parameter estimates (means and covariances) are substituted into $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The term $(\mathbf{y}_i - \boldsymbol{\mu}_i)' \boldsymbol{\Sigma}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ is of particular interest, as it quantifies the discrepancy between a single individual's scores and the parameter values at a particular iteration—readers who are familiar with multivariate statistics may recognize this term as Mahalanobis distance. Summing Equation 2 across the entire sample produces a log likelihood value that quantifies the relative probability that the data were sampled from a normally distributed population with a particular mean and covariance matrix ($\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, respectively). At each iteration, the values of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are adjusted in an attempt to identify the set of values with the highest log likelihood (i.e., pro bability of producing the sample data).

ML estimation is ideally suited for missing-data problems because each person's score vector, $\mathbf{y}_i$, need not be complete. The i subscript in Equation 2 indicates that the elements in each person's data vector may differ in number and content. To illustrate, suppose it were of interest to estimate the means and covariance matrix for three variables: hours spent doing homework (HW), parental involvement in homework (PI), and reading achievement (RA). For students with complete data, the Mahalanobis distance portion of Equation 2 would be computed as

$$\left(\begin{bmatrix} y_{HW} \\ y_{PI} \\ y_{RA} \end{bmatrix} - \begin{bmatrix} \mu_{HW} \\ \mu_{PI} \\ \mu_{RA} \end{bmatrix}\right)' \begin{bmatrix} \sigma^2_{HW} & & \\ \sigma_{HW\text{z}PI} & \sigma^2_{PI} & \\ \sigma_{HW\text{z}RA} & \sigma_{PI\text{z}RA} & \sigma^2_{RA} \end{bmatrix}^{-1} \left(\begin{bmatrix} y_{HW} \\ y_{PI} \\ y_{RA} \end{bmatrix} - \begin{bmatrix} \mu_{HW} \\ \mu_{PI} \\ \mu_{RA} \end{bmatrix}\right).$$

In contrast, students with missing homework effort scores would have their Mahalanobis distance computed on the basis of only the two variables for which they had data, as shown here:

$$\left(\begin{bmatrix} y_{PI} \\ y_{RA} \end{bmatrix} - \begin{bmatrix} \mu_{PI} \\ \mu_{RA} \end{bmatrix}\right)' \begin{bmatrix} \sigma^2_{PI} & \\ \sigma_{PI\text{z}RA} & \sigma^2_{RA} \end{bmatrix}^{-1} \left(\begin{bmatrix} y_{PI} \\ y_{RA} \end{bmatrix} - \begin{bmatrix} \mu_{PI} \\ \mu_{RA} \end{bmatrix}\right)$$

When using ML estimation there is no need to discard cases that have incomplete data, nor is it necessary to "fix" the data (e.g., impute missing values with the mean) before running the analysis; estimation is based on all available data points, and subjects need not have complete data.

It is not obvious from the previous equations, but the inclusion of cases with partial data actually contributes to the estimation of all parameters. Although missing values are not imputed during this process, the partial data do imply probable values for the missing scores via the correlations among the variables. To illustrate, a small educational data set containing the number of absences and achievement scores for a hypothetical sample of 10 students is given in Table 1. To simulate an MAR missing data, achievement scores were deleted for the three students with the highest number of absences (i.e., missing scores on the achievement variable were related to another measured variable). Also, a single absence score was randomly deleted (i.e., an MCAR mechanism). These data are useful strictly for demonstration purposes, and it would normally be unwise to use ML estimation with such a small sample. The means, standard deviations, and correlation between absences and achievement were estimated by using ML, listwise deletion, and mean imputation, and the results of these analyses are given in Table 2.

If the complete data estimates are viewed as the "true" values in the example, it is clear that ML estimation produces estimates that are relatively free of distortion, particularly when compared to listwise deletion and mean imputation. To put the bias in some perspective, the listwise deletion achievement mean is "off" by about one standard error unit, while the distortion in the absence mean is equi-valent to nearly two standard error units. Not only were these results expected on the basis of Rubin's (1976) theoretical work, but it is fairly straightforward to understand why the traditional methods performed poorly. Recall that achievement scores were selectively missing for cases with high absences. Because the two variables were negatively correlated ($r = -0.57$),

TABLE 1
Hypothetical education data

Complete		MAR missing	
Absences	Achievement	Absences	Achievement
0	53	0	53
2	61	2	61
6	70	6	70
7	47	?	47
8	38	8	38
9	53	9	53
11	47	11	47
14	53	14	?
15	43	15	?
18	37	18	?

Note. Question marks indicate that data points were deleted to simulate MCAR and MAR.

TABLE 2
Missing data parameter estimates from hypothetical education data

Variable	Complete data	Missing-data method		
		Maximum likelihood	Listwise deletion	Mean imputation
Mean				
Absences	9.00	9.32	6.00	9.22
Achievement	50.20	50.26	53.67	52.71
Standard deviation				
Absences	5.38	5.56	3.87	5.34
Achievement	9.63	10.15	10.09	8.06
Correlation				
Absences/achievement	−.57	−.50	−.39	−.26

the lower tail of the achievement distribution was truncated, as was the upper tail of the absence distribution—the bias in the listwise deletion mean estimates reflects this. In contrast, ML uses all observed data and incorporates the partial data vectors during estimation. Although achievement scores are missing for three cases, the inclusion of absence data for these students implies different parameter values than would be obtained if these cases were removed from the analysis. Although it may not be obvious from the equations presented earlier, ML essentially "borrows" information from the absence data to estimate the achievement parameters, and does so via the linear relationship between these two variables. As described previously, replacing missing values with the arithmetic mean adds nothing to the numerator of the covariance (and variance) formula, thereby negatively biasing variances and covariances (and thus correlations).

One might reasonably argue that the preceding example was "rigged" in favor of ML estimation, and it is true that scores were deleted in a way that would disadvantage listwise deletion. However, this scenario demonstrates an important finding that follows from Rubin's (1976) theoretical work, namely that ML estimation will provide optimal performance in situations where traditional methods fail. As noted previously, one might view ML estimation as being more "robust" in the sense that it requires less strict assumptions about the missing data, and this example illustrates the point. Again, it is important to note that ML estimation will likely produce biased parameter estimates when data are MNAR (i.e., missing values are related to the underlying values of the variable), and we do not mean to suggest that ML and MI provide a panacea for all missing-data problems.

Before proceeding to MI, we briefly describe the EM (expectation maximization) algorithm, a common method for obtaining ML parameter estimates. EM was originally proposed as a method for obtaining ML estimates of a covariance matrix and mean vector (Dempster, Laird, & Rubin, 1977) but has since been adapted for use in a wide variety of estimation problems (e.g., hierarchical linear

models; Raudenbush & Bryk, 2002). In fact, if one's research questions involve means and covariances (or correlations), EM is a straightforward method for obtaining ML estimates with incomplete data, and the procedure is widely available in software packages.

To be complete, it is necessary to distinguish between two seemingly different applications of EM. The first approach, which might be described as "direct EM," is used to estimate the parameters of a linear statistical model. For example, the direct EM approach is used in conjunction with hierarchical linear models (e.g., HLM, SPSS MIXED procedure, SAS PROC MIXED) and structural equation models (e.g., Mplus, EQS). This application of EM is somewhat different from that outlined by Dempster et al. (1977), because the data set used to fit the model may have no missing values. In this context, EM views the model parameters (rather than the data points themselves) as missing values to be estimated. In contrast, the "traditional" use of EM involves the estimation of a covariance matrix and mean vector from a data set with missing values. We provide a description of EM in this context, given its clear linkage to the goals of this article.

EM is an iterative procedure that repeatedly cycles between two steps: the *E*, or *expectation*, step imputes missing values; and the *M*, or *maximization*, step estimates the covariance matrix and mean vector. To illustrate, we return to the earlier example involving hours spent on homework (HW), parental involvement in homework (PI), and reading achievement (RA).

The first EM step requires an initial estimate of the covariance matrix, which can be obtained by using any number of methods (e.g., listwise deletion). The purpose of the *E* step is to estimate the missing values, and this is accomplished using regression imputation. For example, students with missing homework reports would have their missing values replaced by predicted scores from the regression of HW on PI and RA. In a similar vein, students missing both PI and RA would have their missing values imputed from two separate regressions (PI on HW and RA on HW). As pointed out earlier, a shortcoming of regression imputation is the loss of residual variability present in the hypothetically complete data (the imputed values fall directly on a regression line). To correct this problem, a residual term is added to each imputed value in the *E* step. The purpose of the *M* step is to update the covariance matrix using the "filled-in" data from the previous *E* step. The updated covariance matrix is obtained using standard formulae, and this new covariance matrix is fed into the next *E* step, where new estimates of the missing values are generated, and the two-step process is repeated. The iterative process continues until the difference between covariance matrices from adjacent *M* steps differs by some trivial amount.

The end product from an EM analysis is an ML estimate of a covariance matrix and mean vector (not an imputed data set). Because correlations are closely related to covariances ($r = \text{cov}_{xy}/\sigma_x\sigma_y$), an EM covariance matrix can readily be converted to a matrix of correlations. Although it is more computationally intensive to do so, standard errors can also be obtained from EM, and the resulting covariances can be tested for statistical significance. An EM covariance (or correlation) matrix can be obtained from a variety of different software packages, including SPSS (the Missing Values Analysis procedure), SAS (PROC MI), structural equation modeling packages (e.g., Mplus), and stand-alone freeware packages available on the Internet (e.g., EMCOV and NORM).

Multiple Imputation

As noted previously, one of the problems with imputation procedures is that the "filled-in" data set generally lacks variability that would have been present had the data been complete. Rather than treating a single set of imputed values as "true" estimates of the missing values, MI creates a number of imputed data sets (frequently between 5 and 10), each of which contains a different plausible estimate of the missing values. The analysis of interest is subsequently performed on each of the m imputed data sets, and the parameter estimates (e.g., regression coefficients) are averaged to produce a single set of results. Different MI algorithms have been proposed (Allison, 2000), but we focus on Bayesian MI, popularized by Schafer (1997). Schafer's approach can be implemented in a point-and-click environment using the NORM freeware package and is also available in the SAS MI procedure; SPSS currently offers no MI facility. MI is arguably more complex than ML estimation, and space limitations preclude a complete discussion of the topic. However, a number of accessible introductions to MI are available in the literature (Graham & Hofer, 2000; Schafer & Graham, 2002; Sinharay, Stern, & Russell, 2001); and Horton and Lipsitz (2001) provide an overview of MI software packages.

Applying MI involves three distinct analytic phases: imputation, analysis, and pooling of parameter estimates. The imputation phase is, itself, an iterative procedure involving two steps, *I* and *P*. The *I*, or *imputation,* step resembles EM in that a covariance matrix is used to construct a series of regression equations, and missing values are replaced by the predicted scores from these equations. Residual variation is restored to the imputed data points by adding a random draw from the normal distribution of residuals for a particular variable.

Thus far, MI sounds quite similar to the imputation process used by EM or stochastic regression imputation (i.e., values are imputed and residual variation is restored). What sets MI apart is that *different* estimates of the population covariance matrix are used to create each of the m imputed data sets. MI recognizes that the missing data produce some uncertainty in the covariance matrix used to generate the imputed values. Thus, in the *P,* or *posterior,* step of the imputation phase, new covariance matrix elements are randomly sampled from a distribution of possible values (in Bayesian terminology, known as a *posterior distribution*) based on the filled-in data from the previous *I* step. Thus the imputation phase of MI involves a two-step procedure whereby missing values are imputed via a series of regression equations, and new covariance matrix elements are sampled from a distribution of values that are consistent with the previously imputed data. The new estimate of the covariance matrix is used to construct new regression equations in the next *I* step, and the process is repeated.

The ultimate goal of the imputation phase is to produce m imputed data sets, each of which is "filled in" with values that are essentially a random draw from a distribution of plausible missing values. However, the imputed values in adjacent *I* steps are correlated with one another (a process called *autocorrelation*), so the imputation phase cannot simply be repeated 10 times (if $m = 10$ imputed data sets are desired, for example). In practice, random draws are simulated by allowing a sufficient number of *I* steps to lapse between each retained data set. For example, if it was of interest to generate $m = 10$ imputed data sets, then 1,000 iterations (cycles of *I* and *P* steps) could be performed, and every 100th imputed data set could be retained for future analyses.

Following the creation of the m imputed data sets, the data are analyzed and the parameter estimates are pooled, or averaged. Unlike ML, the analysis phase of MI may be quite distinct from the missing-data handling phase (we avoid the phrase "imputation phase" because ML does not impute missing values). That is, a general imputation strategy might be used with no particular analytic goal in mind, whereas ML estimates a specific analytic model (e.g., a multiple regression, structural equation model). Regardless of whether the imputation phase was performed with a specific analysis in mind, the analysis phase involves fitting the desired statistical model to each of the imputed data sets. Note that no special software is needed in this phase, as the m analyses are being conducted on complete data sets.

The final MI phase involves pooling the parameter estimates and standard errors from the analysis phase. For example, suppose it were of interest to estimate the regression of student achievement on number of absences, as in the example above. Assuming that $m = 10$ imputed data sets were created, the 10 regression coefficients and standard errors would be collected into a new data file and subsequently aggregated. A single estimand would be computed by simply averaging the b_m. In a similar vein, standard errors are aggregated by taking into account the average estimated standard error (technically, squared standard error, or variance) and the variability of the parameter values across the 10 analyses. This process will be demonstrated in the final section of the article.

Some final points should be made about MI. First, the definition of convergence is quite different from that of ML. ML estimation is said to converge when parameter values no longer change from one iteration to the next (they converge to a single set of values). In contrast, MI involves sampling new covariance matrix elements at every cycle of the imputation phase, so parameter values never converge in value. Instead, convergence is attained when the *distribution* of parameters stabilizes. In practice this is achieved by allowing a number of "burn in" cycles (e.g., 200) to lapse before retaining the first imputed data set. We return to this issue in more detail in the final section of the article.

Second, MI also requires the multivariate normality assumption. However, there is some evidence that MI performs well under fairly substantial violations of normality (Graham & Schafer, 1999). On a related issue, Schafer (1997) suggests that nominal and ordinal variables can be used in the imputation process, and MI software packages offer a number of useful options in this regard (e.g., transformations, dummy coding, and rounding imputed values of discrete variables).

Finally, as mentioned previously, MI will produce unbiased parameter estimates when data are MAR but is likely to produce biased estimates when data are MNAR.

Comparison of Modern Missing-Data Techniques

Before proceeding to the methodological review, it is important to point out the similarities and differences between ML and MI. As discussed previously, the primary advantage of both techniques is that they require less strict assumptions about the missing data. This is not to imply that ML and MI will never produce biased results. Quite the contrary—they will be biased if missingness is due to the outcome variable itself (i.e., data are MNAR).

In terms of parameter estimates, ML and MI should produce similar results when the same set of cases and variables is used, although MI standard errors may

be slightly larger (Collins, Schafer, & Kam, 2001; Schafer & Graham, 2002). One of the primary advantages of MI is the ease with which auxiliary variables can be incorporated into the imputation process. For example, one might wish to impute missing values using a superset of the variables included in the ultimate analysis, under the belief that the auxiliary variables may be related to missingness, thereby reducing possible bias. Because the imputation and analysis phases are distinct from one another, auxiliary variables can be used to impute missing values but need not appear in the analytic model. This is not true for ML, however, where missing-data handling is built directly into the estimation of the substantive model.

To illustrate, suppose it is of interest to regress student achievement on a number of educational predictors. In addition, it is suspected that parental socioeconomic status (SES) may be related to missing data in a study (e.g., perhaps parents from disadvantaged homes are less likely to return consent forms), but SES is not a variable that appears in the regression analysis. For MAR to hold, SES scores must be taken into account. This is straightforward for the MI user, as SES can serve as a predictor in the imputation model but need not appear in the substantive regression model, because the imputed values are already conditioned on SES. In contrast, SES must somehow be incorporated into the ML analysis, as missing-data handling and estimation are concurrent processes. Simply adding SES as an additional predictor is not a viable option, as doing so would alter the regression coefficients specified by the substantive research questions. Fortunately, Graham (2003) outlined a method for incorporating auxiliary variables into an ML analysis that uses existing structural equation modeling software. Because structural equation modeling is an increasingly general analytic framework, Graham's (2003) approach can be used for a variety of linear model analyses (e.g., correlation, regression). We will demonstrate the inclusion of auxiliary variables using both ML and MI in the last section of this article.

On a related point, ML and MI differ in their level of generality and thus in the breadth of analyses to which they can be applied. Again, the fact that the imputation and analysis phases are distinct gives MI a distinct advantage in this regard. Because the substantive analysis is performed on (multiple) complete data sets, MI can be used in conjunction with virtually any analytic model. In contrast, ML is model-specific, in the sense that missing-data handling is built into the estimation process. Thus, for ML to be applicable to a particular analysis, an estimation routine must be available in an existing software package. Fortunately, the number of models for which an ML missing-data estimator is available is growing rapidly (e.g., Mplus 3; Muthén & Muthén, 2004).

Methodological Review of Reporting Practices

Having provided an overview of the missing-data concepts, we now present the findings from a methodological review of applied educational and psychological literature. The purpose of this review was to investigate missing-data reporting practices in a sample of applied research journals. In doing so, our primary concern was to document the methods used to treat missing data and the reporting practices used by authors who analyze incomplete data sets.

Roth (1994) reviewed a random sample of 45 articles from the *Journal of Applied Psychology* and *Personnel Psychology* between 1989 and 1991 and found listwise deletion and pairwise deletion to be the predominant approaches employed

by empirical studies. This finding is not surprising given that the recent growth spurt in the missing-data literature (and concurrent implementation of ML and MI in software packages) did not take place until the latter half of the 1990s. Accordingly, we examined empirical articles from two years: 1999 and 2003. The year 1999 represents a demarcation line of sorts: In that year, a report by the APA Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999) specifically discouraged the use of listwise and pairwise deletion, stating that these methods are "among the worst methods available for practical applications" (p. 598). Since the publication of that report, the number of software options for implementing ML and MI has grown to the point where software availability is now no longer a limiting factor.

The data for the 1999 methodological review consisted of research articles published in 16 educational and applied psychological journals. The selected journals represent a variety of disciplines within the field of education and psychology, and they closely mirror those used in a methodological review of ANOVA practices published by Kesselman et al. (1998) in the *Review of Educational Research.* Having found no instances of ML or MI in the 1999 volumes, we expanded the list of 2003 journals to 23. The selected journals and the number of articles examined from each are listed in Table 3. We acknowledge that the list of journals in Table 3 is not a random sample, nor does it reflect the full breadth of disciplines in education and psychology. Although many important journals were omitted from the list, the included journals do represent a broad cross-section of disciplines and are consistent with previous methodological reviews published in the *Review of Educational Research.*

The methodological review excluded a number of writings that were not relevant to the goals of our study, including meta-analyses, qualitative studies, policy or position papers, and literature reviews. The frequencies listed in Table 3 reflect only those articles retained for analysis. Note that the frequencies represent a census of the 1999 articles and a random sample of 50% of the articles in each of the 2003 volumes. Finally, a substantial number of articles contained multiple studies (Study 1, Study 2, etc.). In documenting the prevalence and use of missing-data procedures, we chose to define the study, rather than the article, as the unit of analysis. That is, multiple studies from the same article were treated as separate cases, provided that independent samples were used in each study.

Identifying the presence of, and enumerating the amount of, missing data proved extremely difficult. In a typical study, details concerning missing data were seldom reported, particularly in 1999. In cases where authors did not explicitly acknowledge missing data, we examined the discrepancy between the reported degrees of freedom for a given analysis and the degrees of freedom that one would expect on the basis of the stated sample size and design characteristics (e.g., number of ANOVA factors). A discrepancy in the degrees of freedom values (or degrees of freedom that changed across analyses) thus indicated that data were missing.

The methods used to handle missing data were, in many cases, difficult to ascertain because explicit descriptions of missing-data procedures were rare. The most commonly employed methods were listwise deletion, pairwise deletion, or a combination of the two. Recall that listwise deletion removes cases with missing values, whereas pairwise deletion omits cases on an analysis-by-analysis basis. Although the distinction seems clear enough, these methods were not mutually exclusive within

TABLE 3
Frequency distribution of articles used in methodological review

Journal	1999 Studies	2003 Studies
American Educational Research Journal	7	3
Child Development	105	63
Cognition and Instruction	7	N/A
Contemporary Educational Psychology	14	10
Developmental Psychology	76	34
Educational and Psychological Measurement	N/A	11
Educational Evaluation and Policy Analysis	N/A	5
Educational Technology, Research and Development	7	4
Journal of Applied Psychology	67	36
Journal of Applied Sport Psychology	13	8
Journal of Counseling Psychology	36	21
Journal of Educational Computing Research	9	7
Journal of Educational Psychology	57	31
Journal of Experimental Child Psychology	38	21
Journal of Experimental Education	12	2
Journal of Personality and Social Psychology	143	46
Journal of Research in Mathematics Education	N/A	1
Journal of Research in Science Teaching	N/A	11
Journal of School Psychology	N/A	7
Journal of Special Education	N/A	3
Modern Language Journal	N/A	4
Reading Research Quarterly	11	6
Research in Higher Education	N/A	15
Sociology of Education	8	4

Note. N/A = studies were not sampled from this journal for analysis.

a given study. For example, a number of studies described the blanket removal of cases with missing data (i.e., listwise deletion) but went on to report sample sizes that varied across analyses (i.e., pairwise deletion). Thus we ultimately chose to classify studies as using either "traditional methods" (e.g., listwise deletion, pairwise deletion, mean imputation, or regression imputation) or "modern methods" (e.g., ML or MI).

Because the use of traditional missing-data methods was generally difficult to identify, a set of coding criteria were established. A study was coded as using a traditional missing-data technique if any of the following criteria were met: (a) The author included a statement describing the blanket removal of incomplete cases; (b) a discrepancy was identified between the reported and expected degrees of freedom; (c) the sample size associated with a latent variable model fit statistic (e.g., the χ^2 statistic) was different from the stated sample size in the description of methods; (d) the author included a statement acknowledging that sample sizes varied across analyses; (e) sample sizes varied across table entries (correlation matrices, tables of descriptive statistics, etc.); or (f) the author explicitly acknowledged using a traditional imputation method (e.g., mean or regression imputation). Based on the extra analytic steps or specialized software required to implement ML and MI,

we reasoned that the use of "modern" techniques could be identified only from explicit descriptions of the procedure in the body of text.

Results of 1999 Methodological Review

Prevalence of Missing Data

Of the 989 studies (i.e., independent samples) that we examined, 737 (74.52%) had no detectable missing data, 160 (16.18%) had missing data, and 92 (9.3%) were indeterminate (e.g., degrees of freedom were not reported). Based on our experience and the results of the 2003 review presented below, the 16% prevalence rate may represent a gross underestimate, given that missing data were impossible to detect in many cases. For example, studies using listwise deletion would be identified only if the author explicitly described the procedure. In a similar vein, we suspect that item-level missing data were frequently handled by computing scale scores from the available items. Again, this practice is impossible to detect without an explicit description of the procedure.

Proportion of Missing Data

Considering only those studies that we had identified as having missing data, the proportion of missing cases per analysis ranged from less than 1% to approximately 67%, and the mean percentage of cases with missing data was $M = 7.60$ ($SD = 8.07$). Not surprisingly, we observed differences between cross-sectional and longitudinal analyses. The maximum proportion of missing cases observed in a cross-sectional design was 27.84%, as compared with a maximum of approximately 67% in a longitudinal design. The means for cross-sectional and longitudinal designs were $M = 7.09$ ($SD = 6.07$) and 9.78 ($SD = 13.50$), respectively.

It is important to understand that these missing-data rates represent the proportion of missing cases per analysis, not the proportion of missing scores. To illustrate, consider a multiple regression analysis. A single degree of freedom discrepancy in the error term (the method we used to compute these percentages) could be caused either by a case with a single missing value or by a case with multiple missing values. Therefore, the missing-data rates reported here should be interpreted as representing the proportion of cases excluded from a single analysis, not (a) the proportion of cases in the sample having missing values, or (b) the proportion of missing values in the data matrix. Unfortunately, these latter two quantities are impossible to obtain unless explicitly reported.

Missing-Data Techniques

The 160 studies that we identified as having missing data relied exclusively on traditional missing-data techniques, and we observed no instances of ML estimation or MI. Consistent with Roth (1994), approximately 96% of the articles we reviewed used listwise deletion, pairwise deletion, or some combination of the two. In addition, five studies used some form of mean imputation, and a single study used regression imputation.

Exemplars of Reporting Practice

The fact that no instances of ML or MI were observed is, itself, problematic given the preponderance of recent methodological studies favoring these approaches.

The level of attention given to reporting missing data in the 1999 articles is, perhaps, equally problematic. One anonymous reviewer of this article characterized missing data as a "dirty little secret" of educational research, and we feel that this statement accurately describes the treatment of missing data in the published research reports that we reviewed. To illustrate "typical" practice, we provide a number of exemplars from those articles. In doing so, it was not our intent to single out individual studies for the purpose of imposing a value judgment. Rather, it was our hope that these examples might provide an impetus for improving reporting practices. In the spirit of this goal, we take an unorthodox approach and exclude citations from the quoted material presented in this section in an effort to protect the anonymity of the original works. Detailed citations will be made available to readers upon request.

Consistent with Roth (1994), it was relatively rare for authors to explicitly mention missing data in their research reports. Considering the 160 studies that we identified as having missing data, only 54 (33.75%) explicitly acknowledged the problem. In many cases we hesitate to characterize these reports as "explicit," because discussions of missing data were relegated to a footnote or table note. In contrast, 106 (66.25%) of the studies that we identified as having missing data did not mention the problem, and missing values were inferred from degrees of freedom values that were inconsistent with the stated sample size and design characteristics.

In cases where authors did explicitly mention missing data, it was extremely rare to name the technique used. For example, we found only three articles that used the term "listwise" to describe their method, and only one used that term in the body of the article (as opposed to a table note). These authors stated, "The listwise deletion sample of families with no missing data produced an n of 190 for the analysis." It was common for authors to use words such as "eliminated," "excluded," and "discarded" to describe a listwise reduction in the overall sample size. For example, one study reported, "Of the 199 children in this sample, 194 had complete data on all measures and were included in the analysis (5 had missing data on one or more measures and were excluded)." Similarly, another study reported, "Another 13 individuals completed surveys but were not counted as respondents because their surveys were either missing too much data or the responses appeared to be confused."

In a similar vein, we found only a single article that used the word "pairwise" to describe the method used, and that reference appeared in a table note. Among studies that explicitly mentioned missing data, the most common strategy used to convey pairwise deletion was a statement highlighting sample size differences across analyses (e.g., "Sample sizes may vary slightly from one task to the other because some participants were absent on certain occasions and tasks could not be readministered"). Similarly, one author explained the discrepancy in the reported degrees of freedom in a footnote as follows: "Degrees of freedom vary slightly across analyses because of occasional missing data."

Given that the majority of studies made no mention of missing data, it is probably not surprising that only three articles made an attempt to test the MCAR assumption. In one case, the authors created a dichotomous dummy variable that denoted whether a subject had missing data at the final measurement occasion of a longitudinal study, and compared these two groups on a number of response variables (equivalence of means would lend some support to the MCAR assumption). In a

footnote, the authors stated, "*T*-tests were used to compare study participants who had parental reactions measures at T5 with those who did not (i.e., had dropped from the sample or had missing data on the target parent reactions measure) on all of the major variables used in the analyses." The authors go on to state that no significant differences were observed.

Finally, we found situations where authors justified their handling of missing data on the basis of inaccurate information. For example, one author explained, "Due to concerns about these sample biases, and the percentage of cases that would be deleted using listwise deletion, mean substitution was made for those cases missing moderate amounts of data." This author went on to report that 14% of the cases had their missing values imputed. Interestingly, the author's rationale for employing mean imputation was based on concerns about parameter estimate bias, even though it is known that mean imputation produces substantial bias of virtually every parameter except the mean (e.g., Little & Rubin, 1987). In another case, the study's author used a series of *t* tests to compare observations that had missing values with those that had complete data. After finding significant differences, the author incorrectly attempted to remedy the situation with mean imputation. He stated that "[a]fter mean substitution . . . there were fewer significant differences between groups." Unfortunately, this statement implies that bias due to missing data was somehow eliminated with mean imputation, when it is more likely that the lack of group differences resulted from an infusion of bias due to mean imputation.

Results of 2003 Methodological Review

The articles reviewed thus far were published in the same year as the report from the APA Taskforce on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999), which encouraged authors to report complications such as missing data and discouraged the use of listwise and pairwise deletion. Although the 1999 articles serve as a useful baseline for comparison, it is unreasonable to expect the taskforce's recommendations to be evident in these research reports. Thus we reviewed the 2003 volumes of 23 journals in an attempt to document changes in missing-data reporting practices that have occurred since the 1999 report.

Of the 545 studies examined, 288 (52.84%) had no detectable missing data, 229 (42.02%) had missing data, and 28 (5.14%) were indeterminate (e.g., degrees of freedom were not reported). What is striking about these findings is that the proportion of studies with missing data increased from 16% in 1999 to 42% in 2003. Although this might appear odd, the apparent increase in the prevalence of missing data can be attributed to a change in reporting practices. As noted earlier, we were unable to detect the presence of missing data in many cases (e.g., listwise deletion) unless authors explicitly acknowledged the problem. In 1999, it was relatively rare for authors to acknowledge missing data; only 33.75% of the studies that we identified as having missing data explicitly acknowledged the problem. However, the proportion of studies that acknowledged missing data increased dramatically in 2003; of the 229 studies that we identified as having missing data, 170 (74.24%) explicitly mentioned the problem. Thus it appears that more studies are adhering to this recommendation by the APA taskforce: "Before presenting results, report complications, protocol violations, and other unanticipated events in data collection. These include missing data, attrition, and nonresponse" (Wilkinson & Task Force on Statistical Inference, 1999, p. 597).

Another positive, albeit minor, shift in reporting practices has to do with testing the MCAR assumption. Only 3 studies in the 1999 sample examined mean differences between respondents and nonrespondents (support for MCAR is established if respondents and nonrespondents have equal means on other study variables), but we found 15 studies in the 2003 sample that did so. Although this is a relatively small percentage of the 229 studies that we identified as having missing data (6.5%), it nevertheless represents an improvement. Note also that the level of reporting detail varied across studies; some studies mentioned these tests in a single sentence, while others devoted considerable space to the issue.

Although the increase in the proportion of studies that explicitly mentioned missing data was a major improvement in reporting practices, the methods used to analyze the data changed very little in 2003. Of the 229 studies that we identified as having missing data, we found only 6 that used ML estimation or MI. Fifteen additional studies used growth modeling techniques to analyze repeated measures data, but the missing-data technique used by these studies was unclear. It is quite likely that many of these studies used ML estimation, given that it is the default missing-data handling procedure in some growth modeling software packages. However, this assumption is not necessarily safe, as we found growth modeling studies that performed listwise deletion in spite of the software defaults.

In the interest of saving space, we do not provide exemplars of problematic reporting practices observed in the 2003 articles; the 1999 results are quite representative in this regard. Instead, we highlight 2003 articles that used ML estimation and MI, as these articles may serve as useful models for researchers preparing their manuscripts.

McQueen, Getz, and Bray (2003) might be viewed as a model article, at least with respect to missing-data reporting practices. These authors devoted three paragraphs in the methods section to their exploration and handling of missing data and provided the following description of ML estimation:

> Amos 4.0 software was used to test the hypothesized models in this study. Amos does not impute missing values but instead recognizes missing data and uses all observed data values to estimate models with a full information maximum likelihood (Anderson, 1957) approach. With this sample, Amos identified 2,441 cases containing some or all of the data being analyzed. Full information maximum likelihood is the recommended estimation method of choice when the data are missing at random, and it may be less biased than other multivariate approaches when missing data are nonignorable. (p. 1741)

In a similar vein, Snyder et al. (2003, p. 1884) and Sadler and Woody (2003, p. 84) provided explicit descriptions of ML estimation for missing data in their methods sections.

Only two studies, Hill, Brooks-Gunn, and Waldfogel (2003) and Tolan, Gorman-Smith, and Henry (2003), used MI to treat missing values before the analysis. Hill et al. (2003) offered the following description of their procedure:

> We used multiple imputation (MI; Baer, Kivlahan, Blume, McKnight, & Marlatt, 2001; Rubin, 1987; Schafer, 1997), which replaces missing values with predictions based on all the other information observed in the study. MI relies on more plausible assumption than do standard approaches (listwise deletion or complete case analysis), properly accounts for our uncertainty

about the missing values (leading to appropriate standard errors), and retains the original sample size of the study. (p. 735)

Illustrative Analyses

If the routine use of ML and MI represents one of the fundamental "changes in research over the next decade" (Azar, 2002, p. 70), it is important for applied researchers to gain some exposure to these new analytic methods. In this section we illustrate ML estimation and MI using a sample of 3,116 cases extracted from the Longitudinal Study of American Youth (LSAY).

To keep the analytic model simple, ninth-grade math composite scores were regressed on four predictor variables measured in the seventh grade: minority status (0 = Caucasian, 1 = Minority); educational expectations (coded on a 6-point scale anchored at "high school only" and "Ph.D."); parental academic encouragement; and peer academic encouragement (the two being scale scores expressed on the *z*-score metric). Although not in the regression model, three additional variables served as "auxiliary variables": average parental socioeconomic index, home resources, and a dichotomous problem behavior indicator (0 = no problem behavior, 1 = the student was expelled or arrested, or dropped out). The variables used in this example were not intended to test meaningful substantive hypotheses but were chosen to illustrate certain nuances of ML estimation and MI.

A brief discussion of auxiliary variables is warranted at this point. Auxiliary variables can be thought of as variables unrelated to one's hypotheses but possibly related to the propensity for missing data. The inclusion of auxiliary variables can play an important role in estimating a model with missing data. To understand why this is so, recall that ML estimation and MI rely on the untestable assumption that missing values are related to other variables in the data (i.e., the MAR mechanism). To illustrate, suppose that missing ninth-grade math scores are related to parental socioeconomic status (e.g., perhaps families in a particular income bracket have higher mobility rates). If this were true, we would expect regression coefficients from the illustrative analysis to contain some bias because the MAR assumption has not been satisfied (socioeconomic status does not appear in the regression model). However, this bias could be reduced if information from socioeconomic status scores were used during estimation. As we will illustrate, this is quite straightforward with MI, as the imputed data sets are created by using a superset of the variables included in the ultimate analysis. The inclusion of auxiliary variables is not as straightforward with ML estimation but can be readily accomplished with existing software.

We begin the illustrative analyses with some basic data screening. Univariate skewness values for the continuous variables ranged between 0.10 and 0.86 in absolute value, while kurtosis values ranged between 0.16 and 0.68 in absolute value. These values suggest that the data are relatively symmetric, but univariate normality does not necessarily imply multivariate normality (Henson, 1999). However, Mardia's (1974) test of multivariate kurtosis (i.e., $b2$, p) was not statistically significant, $MK = 47.57$, $z = -0.95$, $p = .34$, which indicated that the multivariate normality assumption was satisfied. Four of the variables (educational expectations, the problem behavior indicator, parental academic encouragement, and peer academic encouragement) had very few missing data—the largest missing-data

rate was 1.28%. Missing data were more problematic for the remaining variables: minority status (5.07%), home resources (8.22%), average socioeconomic index (12.52%), and ninth-grade math scores (28.8%).

Although the proportion of missing data is important (and rarely reported), the missing-data mechanism bears heavily on the issue of parameter estimate bias. As noted previously, it is only possible to test for an MCAR mechanism, as any test of MAR or MNAR would require knowledge of the missing values. One method for examining whether data are MCAR is to create dummy codes for each variable with missing values (e.g., 0 = nonrespondents, 1 = respondents) and compare the group means of other measured variables (equality of group means suggests that respondents and nonrespondents do not systematically differ on some other measured variable, thus providing support for MCAR). Clearly, this procedure can produce a substantial number of tests, so multiple comparison issues are certainly relevant. As always, significance tests of these group differences should be accompanied by effect size estimates (Thompson, 1999). To illustrate, we created missing-data indicators for minority status and ninth-grade math scores (the two regression variables with substantial missing data), and performed group mean comparisons using the remaining variables in the regression model. Cohen's *d* effect size values for these comparisons ranged between 0.05 and 0.28 (the average *d* was 0.16), suggesting minimal deviations from MCAR.

Little (1988) proposed a multivariate test statistic for MCAR that is akin to a simultaneous test of all mean differences described above. In fact, Little's procedure reduces to a simple *t* test when data are bivariate and missingness is restricted to a single variable. Little's (1998) test is available in commercial software packages (e.g., SPSS Missing Values Analysis), and a custom SAS program for implementing the test can be downloaded at *http://manila.unl.edu/ckenders*. For the LSAY example, Little's MCAR test was statistically significant, $\chi^2(31, N = 3{,}116) = 111.26, p < .01$, which suggests that the five LSAY variables cannot be described as MCAR. Although this test appears to be at odds with the previous *d* values, one must consider the fact that the chi-square test is very powerful with a sample this large, and is quite capable of detecting relatively small mean differences between respondents and nonrespondents.

After screening the LSAY data, one might reasonably conclude that listwise deletion will perform adequately, given the apparent minor deviations from MCAR. Although the listwise deletion sample size ($n = 2{,}134$) is likely to provide sufficient power in this example, the removal of nearly 1,000 cases from the data is undesirable and unnecessary. Thus we proceed by demonstrating ML and MI in the context of the LSAY data.

Maximum Likelihood Estimation

Structural equation modeling (SEM) software provides a convenient platform for implementing ML estimation because (a) many common analytic models (e.g., regression, ANOVA) are special cases of structural equation models; and (b) all commercial SEM packages offer ML estimation with missing data. The LSAY regression model is expressed in path diagram form in Figure 1.

Note that the regression model in Figure 1 does not include the three auxiliary variables. Graham (2003) proposed two methods for incorporating auxiliary variables into a ML analysis, and we illustrate one of these approaches, the so-

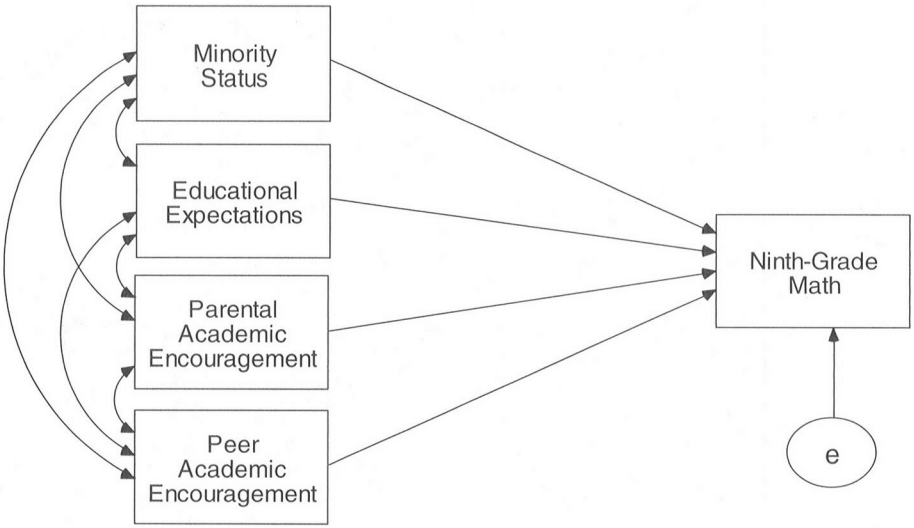

FIGURE 1. *LSAY regression model expressed as a path diagram. Regression coefficients are denoted by straight arrows, correlations by curved arrows. The rectangles represent the five observed variables. The residual portion of each person's outcome variable score is considered to be an unobserved, or latent, variable, and is represented by the ellipse (e).*

called "saturated correlates model." Graham (2003) outlined three rules that dictate the inclusion of auxiliary variables in a model: The auxiliary variables should be (a) correlated with one another; (b) correlated with observed (not latent) predictor variables; and (c) correlated with the residuals of any observed (again, not latent) outcome variable. Applying these rules to the LSAY example results in the path diagram in Figure 2. To reduce visual clutter, a single rectangle is used to denote the set of auxiliary variables in this figure. Each auxiliary variable would be required to correlate with the predictor variables and residual term (i.e., the curved arrows in the diagram), but would also correlate with other auxiliary variables (this would be denoted by curved arrows among the auxiliary variables in the path diagram). At a conceptual level, it is easy to see that information from the auxiliary variables is incorporated into the model via the correlations with the variables appearing in the regression, but the inclusion of the auxiliary variables does not alter the meaning of the model's substantive parameters (e.g., each regression coefficient partials out the remaining three predictors, not the auxiliary variables).

The model in Figure 2 was estimated by using Mplus 3.01. The parameter estimates and standard errors that resulted are presented in Table 4 (labeled *b* and *SE*, respectively). Although these estimates were derived by means of ML estimation, the interpretation of the regression coefficients (and all other model parameters) is identical to the least squares case that most researchers are familiar with. For example, a unit increase in seventh-grade expectations is expected to result in a

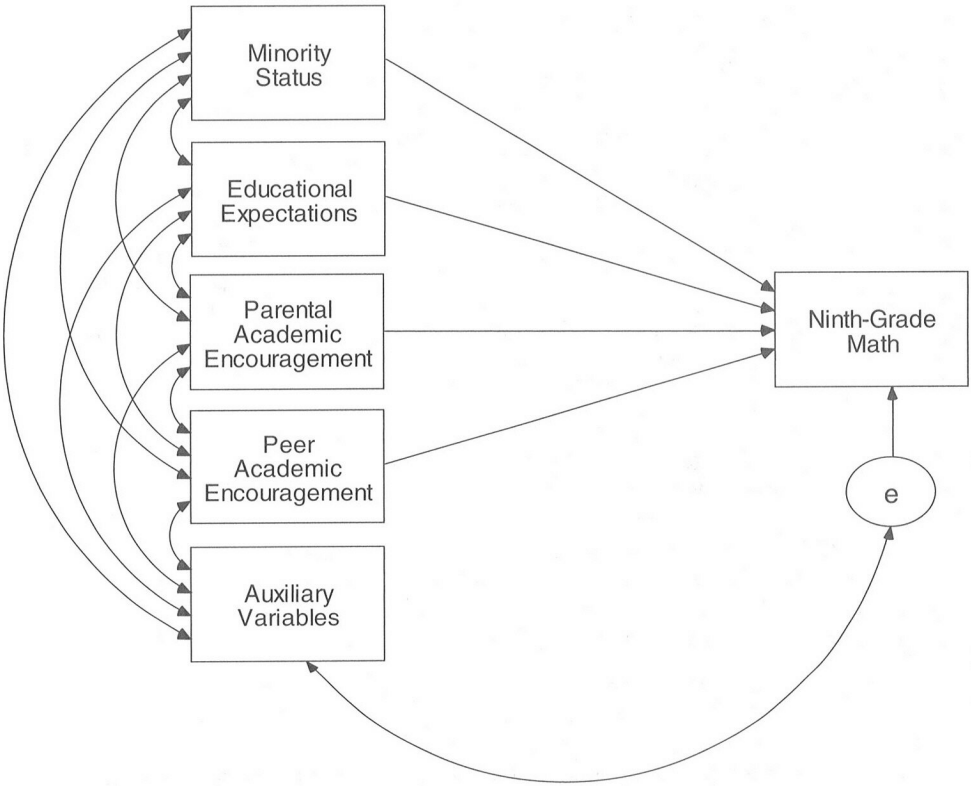

FIGURE 2. *LSAY regression model with auxiliary variables. Regression coefficients are denoted by straight arrows, correlations by curved arrows. The rectangles at left represent the observed and auxiliary variables. The residual portion of each person's outcome variable score is considered to be an unobserved, or latent, variable, and is represented by the ellipse (e).*

3.39 increase in ninth-grade math scores, holding minority status, parental academic encouragement, and peer academic encouragement constant. Consistent with least squares estimation, ML estimation also provides standard error estimates for each parameter. For example, the standard error for the expectations regression coefficient is 0.18; it represents the amount of uncertainty, or sampling error, associated with the coefficient. As is usually the case, a t ratio can be computed by dividing the parameter estimate by its standard error. In the case of educational expectations, $t = 19.16$, which is statistically significant at $p < .05$.

As a technical aside, some software packages offer the option to estimate standard errors on the basis of either the observed or the expected information matrix. When given the option, standard errors should be estimated by using the observed information matrix option, as the resulting standard errors are more appropriate when data are MAR (Kenward & Molenberghs, 1998).

TABLE 4
Results of LSAY regression analysis

Predictor	*b*	*SE*	*t*
Maximum likelihood			
Minority status	−5.94	.56	−10.59
Educational expectations	3.39	.18	19.16
Parental academic encouragement	.73	.27	2.73
Peer academic encouragement	.31	.26	1.19
Multiple imputation			
Minority status	−5.82	.57	−10.22
Educational expectations	3.38	.19	18.25
Parental academic encouragement	.81	.24	3.31
Peer academic encouragement	.26	.27	.99

Note. b = unstandardized regression coefficient estimate.

Multiple Imputation

Unlike the previous ML analysis, MI handles missing data in an imputation phase that precedes the analytic phase. Thus it is first necessary to decide which variables should be used in the imputation model. Schafer (1997, p. 143) suggests that the imputation model should (a) include variables related to the variables being imputed; (b) include variables potentially related to the missingness; and (c) be at least as general as the analysis model (e.g., if the analysis includes an interaction, the imputation model should as well). In the spirit of these guidelines, the LSAY data were imputed using a superset of eight variables that included the five variables from the multiple regression model and the three auxiliary variables.

The MI analysis was carried out with Schafer's (1999) NORM program, which can be downloaded free of charge at the Penn State Methodology Center's website (*http://methodology.psu.edu/downloads/mde.html*). NORM uses a graphical interface that requires no programming or syntax. ASCII (i.e., text) data files are easily imported, and the program offers the user a number of pre- and post-imputation options, such as estimating an EM covariance matrix and mean vector, applying transformations, automatically dummy-coding categorical variables, specifying rounding precision for discrete variables, and pooling MI parameter estimates, to name a few. A more detailed overview of the NORM program can be found in Schafer and Olsen (1998).

The LSAY data were converted to an ASCII file and imported into NORM, which reads data in free format. So a missing value code of −9 was assigned to all variables (NORM assumes a code of −9 by default, but any value can be used). NORM offers the user a number of options for exploring the data before imputation. For example, histograms and normal quantile plots can be used to examine distribution shape. An examination of these plots revealed that the numeric LSAY variables were relatively symmetrical, so we chose not to transform variables before imputation (NORM can automatically implement a number of different power transformations).

Another option that one must consider is the rounding of imputed values. We purposefully included a number of discrete variables in this analysis (e.g., minority

status, home resources) to illustrate this issue. NORM allows the user to (a) round imputed values to a specified decimal or integer; (b) round values to the nearest observed value; or (c) not round at all. For example, consider the home resources variable, for which the observed values were integers ranging between zero and 6. If one chooses not to round, the imputed home resources scores will not be integers, and some values may fall out of range (e.g., < 0 or > 6). To avoid this outcome, we chose to round all discrete variables to the nearest observed value, thus ensuring that the imputed values were consistent with each variable's original metric.

Before creating the imputed data sets, it is usually good practice to estimate the covariance matrix and mean vector by using the EM algorithm, as this can provide insight into the convergence behavior of the MI algorithm (Schafer, 1997). Recall that each imputed data set should simulate a random draw from the distribution of plausible values for the missing data, and this is accomplished by allowing a sufficient number of *I* (imputation) steps to lapse between each imputed data set. To this end, it is frequently suggested that the number of *I* steps separating each imputed data set be at least twice as large as the number of iterations required by EM to converge (i.e., the "2 times the number of EM iterations" rule). In the LSAY example, EM converged after only 14 iterations, which suggests that the MI algorithm might converge even more quickly (Schafer, 1997).

The convergence behavior of the MI algorithm can be assessed more formally by examining time series and autocorrelation function plots, both of which are available in NORM. To explore this issue, we specified 200 iterations of the MI algorithm and saved the variable means and covariances drawn from the posterior distribution at each of the 200 *P* steps (this is accomplished by using options found on the tab labeled "Data augmentation"). Note that the goal of this procedure is to examine the convergence behavior of the MI procedure in this particular set of data (a quality control check); therefore, no imputed data sets were actually saved at this point.

Time series plots display the value of a given parameter (e.g., a mean or covariance) at each *P* step in the imputation process. Time series plots should be examined for all means and covariances, but we illustrate such a plot using the worst linear function (WLF) of the parameters, a scalar function of the means and covariances that converged most slowly (Schafer, 1997, pp. 129–131). The time series plot produced by the NORM program and shown in the upper panel of Figure 3 suggests that the MI algorithm converges reliably, because the values of the WLF are contained within a horizontal band that does not vertically drift or wander across the 200 iterations.

Again, recall that each imputed data set should simulate a random draw from the distribution of plausible values for the missing data. The degree of serial dependence in parameter values can be assessed graphically by using autocorrelation function plots. For a given parameter (e.g., a mean or covariance), the lag-*k* autocorrelation is the Pearson correlation between parameter values separated by *k* iterations. For example, if 100 cycles of the MI algorithm are specified, the lag-3 autocorrelation for the parental encouragement mean is obtained by correlating mean values from iterations 1 to 97 with those from iterations 4 to 100 (i.e., the correlation between mean estimates separated by 3 iterations). Ideally, the autocorrelation should drop to within sampling error of zero within a few iterations (lags), and this is the case with the LSAY data. As can be seen in the lower panel

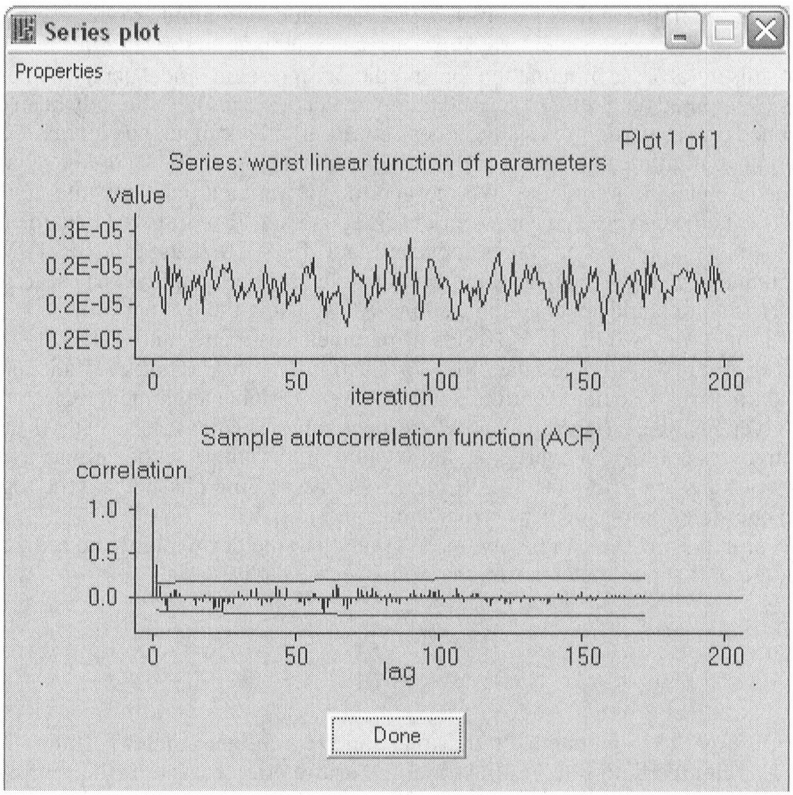

FIGURE 3. *Time series and autocorrelation function plot from LSAY multiple imputation analysis.*

of Figure 3, the autocorrelations for the WLF quickly drop to within sampling error of zero (the confidence interval around zero is shown by the two horizontal lines), which suggests that the serial dependence between imputed data sets dies off rather quickly. A more detailed discussion of MI convergence is found in Schafer (1997, p. 118), along with further examples of time series and autocorrelation function plots.

Having explored the convergence behavior of the MI algorithm, we created $m = 10$ imputed data sets using a single sequential imputation chain; Schafer (1997) recommends the use of 5 to 10 imputed data sets in most cases. Although the EM analysis and autocorrelation function plots suggested the need for relatively few iterations, we opted for a more conservative approach whereby an imputed data set is saved after every 100th cycle of the MI algorithm, beginning with the 200th iteration. In NORM, this is accomplished by setting the number of iterations to 1,100, then specifying that an imputed data set be saved after every 100th iteration (both options are specified in the "Data augmentation" tab). This procedure produces 11 imputed data sets, but only the final 10 sets were subsequently analyzed (i.e., a "burn-in" period of 200 iterations was used).

Once the imputed data sets have been generated, any number of analyses can be conducted by using standard complete-data methods. In the case of the LSAY data, this involved estimating the four-predictor regression model using each of the $m = 10$ imputed data sets. Because the imputed data values have already been conditioned on the auxiliary variables, these three variables can now be ignored. Conducting 10 multiple regression analyses may, at first glance, appear tedious, but the procedure can be automated. We merged the 10 imputed data sets into a single SPSS data file and created a new variable that indexed each imputed data file (i.e., the index variable took on values between 1 and 10). We then used the SPLIT FILE command to generate a separate regression model for each imputed data set (a similar procedure could be accomplished in SAS by using the BY option).

At this point, we had $m = 10$ sets of parameter estimates and standard errors, one set from each imputed data file. To illustrate, the regression coefficients and standard errors associated with seventh-grade educational attainment expectations are given in Table 5 (labeled *b* and *SE*, respectively). The final phase in a MI analysis involves pooling, or averaging, these values into a single set of estimates using rules outlined by Rubin (1987). We illustrate the pooling phase using the regression coefficients and standard errors shown in Table 5.

A single estimand can be obtained for any parameter by taking the arithmetic average of that parameter across the *m* analyses. That is,

$$\overline{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i, \tag{3}$$

where m is the number of imputations and $\hat{Q}_i$ is the parameter estimate from the ith imputed data set. To illustrate, the arithmetic average of the regression coefficients given in Table 5 is $b = 3.38$. As with the ML analysis, the interpretation of this coefficient is identical to the interpretation one would make had there been no missing data; a unit increase in seventh-grade educational expectations should result in a 3.38 increase in ninth-grade math scores, holding other predictors constant.

TABLE 5
Expectations regression coefficients from 10 imputed data sets

Imputation	*b*	*SE*	Variance (SE^2)
1	3.233	.149	.022
2	3.431	.151	.023
3	3.348	.105	.011
4	3.548	.150	.023
5	3.468	.148	.022
6	3.234	.150	.023
7	3.339	.150	.023
8	3.505	.152	.023
9	3.303	.149	.022
10	3.390	.151	.023

Note. *b* = unstandardized regression coefficient estimate.

The rule for combining standard errors is slightly more complicated than combining point estimates. The pooled standard error is a function of two quantities, the within-imputation variance and the between-imputation variance. The within-imputation variance is the arithmetic average of the squared standard errors across the m analyses, or

$$\overline{U} = \frac{1}{m}\sum_{i-1}^{m}\hat{U}_i, \tag{4}$$

where $\hat{U}_i$ is the variance estimate from the ith imputed data set, and m is the number of imputations. From Table 5, the arithmetic average of the m=10 variance estimates (i.e., squared standard errors) is 0.0214.

The between-imputation variance quantifies the variability of the m parameter estimates around the mean estimate, and is given by

$$B = \frac{1}{m}\sum_{i-1}^{m}(\hat{Q}_i - \overline{Q})^2, \tag{5}$$

where m is the number of imputations, $\hat{Q}_i$ is the parameter estimate from the ith imputed data, and $\overline{Q}$ is the mean parameter. Using the coefficients found in Table 5, the between-imputation variance is 0.011.

Finally, the total variance, the square root of which is the pooled standard error, is a function of the within- and between-variance estimates as follows:

$$T = \overline{U} + \left(1 + \frac{1}{m}\right)B. \tag{6}$$

Using the within- and between-imputation variance components calculated previously, $T = 0.033$. Thus, the MI standard error estimate for the expectations regression coefficient is $\sqrt{T}$, or 0.19. Consistent with the pooled parameter estimate, the interpretation of this standard error is identical to the least squares cases. Therefore, 0.19 represents the amount of uncertainty, or sampling error, associated with the educational expectations regression coefficient. As is usually the case, a t ratio can be computed by dividing the parameter estimate by its standard error. In the case of educational expectations, $t = 18.25$, a result that is statistically significant at $p < .05$.

Although the previous computations are useful for pedagogical purposes, parameter estimates and standard errors can be quickly combined using NORM (choose "MI Inference: Scalar" from the Analyze pull-down menu). The parameter estimates and standard errors are highlighted and copied from the SPSS output file and pasted into an ASCII file so that the results from each imputed data set are stacked on top of the next. For example, the LSAY regression results were pasted into a new ASCII file consisting of 40 rows (4 regression coefficients for each of the 10 imputations) and two columns (one for the estimates, the other for standard errors). NORM reports the pooled parameter estimates and standard errors shown above, and also provides t ratios, p values, and 95% confidence intervals for each parameter. The parameter estimates, standard errors, and t ratios for

each predictor variable are given in Table 4. Note that the MI estimates are quite similar to those obtained from the ML analysis in this particular example.

The usual omnibus test of R^2 in multiple regression is akin to simultaneously testing whether all regression coefficients differ from zero (Pedhazur, 1997). Following procedures outlined by Rubin (1987) and by Li, Raghunathan, and Rubin (1991), an approximate F test can be used to test whether the vector of regression coefficients, i.e., $\theta' = (-5.82, 3.38, 0.81, 0.26)$, obtained from the MI analysis is equal to zero (i.e., multiparameter inference). This test requires the strong assumption that the missing information (a function of the missing-data rate and magnitude of relationships among the variables) is the same for all variables, but a simulation study conducted by Li et al. (1991) suggested that accurate significance tests can be obtained even when this assumption is violated, although the test tends to become slightly conservative in that case.

Multiparameter inference is also available in NORM but requires parameter estimates and the parameter covariance matrix from each of the m analyses. The parameter covariance matrix, the diagonal of which contains the squared standard errors, can be obtained from the SPSS REGRESSION procedure by clicking on the "Statistics" button, then selecting "Covariance matrix" from the section of the interface labeled "Regression Coefficients." Note that this covariance matrix contains information about the variance and covariance of the regression coefficients; it is not the same as the covariance matrix of the scores. Consistent with the previous description, the parameter estimates and parameter covariance matrix from each of the m analyses are highlighted and copied from the SPSS output file and stacked in an ASCII file (the NORM Help menu shows the exact layout for this file), and "MI Inference: Multiparameter" is selected from the Analyze pull-down menu. The LSAY data produced a statistically significant F test in this case, $F(4, 120) = 101.61$, $p < .001$, meaning that the set of predictors is significantly different from zero. It is important to note that the derivation of the multiparameter significance test relies on a large N, and the utility of this test in small samples is unclear.

Discussion

During the last 25 years, substantial progress has been made in the area of missing-data analyses. Software for carrying out ML estimation and MI is now widely available, and empirical studies have, almost unequivocally, demonstrated the superiority of these methods over traditional methods such as listwise and pairwise deletion (e.g., Arbuckle, 1996; Enders, 2001a, 2001b, 2003; Enders & Bandalos, 2001; Gold & Bentler, 2000; Graham, Hofer, & MacKinnon, 1996; Graham & Schafer, 1999; Muthén, Kaplan, & Hollis, 1987; Kaplan, 1995; Wothke, 2000).

Although ML and MI may be familiar to many methodologists, one of the goals of this article was to disseminate information about these "modern" missing-data procedures to a wide group of applied researchers. A second goal was to examine the current missing-data analytic practices in a sample of educational and applied psychological journals. As is noted by Keselman et al. (1998), "One consistent finding of methodological research reviews is that a substantial gap often exists between the inferential methods that are recommended in the statistical research literature and those techniques that are actually adopted by applied researchers" (p. 351). Our review of 1999 and 2003 suggests that this statement is definitely true of procedures for handling missing data.

The year 1999 represents a demarcation line of sorts, as a report by the APA Task Force on Statistical Inference (Wilkinson & Task Force on Statistical Inference, 1999) encouraged authors to report complications such as missing data and discouraged the use of listwise and pairwise deletion. The results of our methodological review indicated that authors are reporting their missing data with increased frequency: In 1999, 33.75% of the studies that we identified as having missing data explicitly reported the problem, whereas this number more than doubled, to 74.24%, in 2003. Whether this increase represents changing editorial policies, an increased awareness of missing-data issues, or both, is unclear. Nevertheless, an increase in the proportion of studies that explicitly mention missing data is a clear improvement, and it is hoped that journal editors will continue to encourage sound reporting practices.

However, our methodological review indicated that the methods used to treat missing data have not changed; we identified only six studies in 2003 that definitively used ML or MI. This is probably not surprising, given that major analytic trends—particularly those that are as entrenched as missing-data handling—tend to move somewhat slowly. In addition, software availability almost certainly governs changes in analytic trends. To date, general ML estimation algorithms are available primarily in SEM software packages. Although common classical analyses such as ANOVA and regression can be performed by using SEM software, it is reasonable to expect the use of "modern" missing-data methods to grow slowly until these methods become readily available in statistical packages such as SPSS.

In sum, we hope that applied researchers will consider the biasing impact that missing data can have on their results and take seriously the recommendations put forth by the APA taskforce. In a similar vein, we encourage the adoption of editorial policies that require authors to examine requisite missing-data assumptions (i.e., MCAR) and to implement "modern" missing-data techniques in cases where traditional techniques cannot be justified. The final section of this article presented a heuristic analysis of the LSAY data, and it is hoped that this illustration will serve as a model for researchers who wish to use ML or MI approaches to handle their missing data.

References

Allison, P. D. (2002). *Missing data.* Thousand Oaks, CA: Sage.

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

Arbuckle, J. L. (1996). *Full information estimation in the presence of incomplete data.* Mahwah, NJ: Lawrence Erlbaum.

Azar, B. (2002). Finding a solution for missing data. *Monitor on Psychology, 33,* 70.

Collins, L. M., Schafer, J. L., & Kam, C. M. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods, 6*(4), 330–351.

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, Ser. B, 39,* 1–38.

Enders, C. K. (in press). Estimation by maximum likelihood. In B. Everitt & D. C. Howell (Eds.), *Encyclopedia of behavioral statistics.* West Sussex, UK: Wiley.

Enders, C. K. (2001a). The impact of nonnormality on full information maximum likelihood estimation for structural equation models with missing data. *Psychological Methods, 6,* 352–370.

Enders, C. K. (2001b). The performance of the full information maximum likelihood estimator in multiple regression models with missing data. *Educational and Psychological Measurement, 61,* 713–740.

Enders, C. K. (2003). Using the EM algorithm to estimate coefficient alpha for scales with item level missing data. *Psychological Methods, 8,* 322–337.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 8,* 430–457.

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

Graham, J. W. (2003). Adding missing-data relevant variables to FIML-based structural equation models. *Structural Equation Modeling: A Multidisciplinary Journal, 10,* 80–100.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little & K. U. Schnabel (Eds.), *Modeling longitudinal and multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201–218, 269–281). Mahwah, NJ: Lawrence Erlbaum.

Graham, J. W., Hofer, S. M., & MacKinnon, D. P. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197–218.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In Rick H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 1–29). Thousand Oaks, CA: Sage.

Graham, J. W., Taylor, B. J., & Cumsille, P. E. (2001). Planned missing-data designs in analysis of change. In L. M. Collins & A. G. Sayer (Eds.), *New methods for the analysis of change* (pp. 335–353). Washington, DC: American Psychological Association.

Henson, R. K. (1999). Multivariate normality: What is it and how is it assessed? In B. Thompson (Ed.), *Advances in social science methodology* (Vol. 5, pp. 193–212). Greenwich, CT: JAI Press.

Hill, J. L., Brooks-Gunn, J., & Waldfogel, J. (2003). Sustained effects of high participation in an early intervention for low-birth-weight premature infants. *Developmental Psychology, 39,* 730–744.

Horton, N. J., & Lipsitz, S. R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician, 55,* 244–254.

Kaplan, D. (1995). The impact of BIB spiraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics, 20*(1), 69–82.

Kenward, M. G., & Molenberghs, G. (1998). Likelihood based frequentist inference when data are missing at random. *Statistical Science, 13,* 236–247.

Keselman, H. J., Huberty, C. J., Lix, L. M., Olejnik, S., Cribbie, R. A., Donohue, B., et al. (1998). Statistical practices of educational researchers: An analysis of their ANOVA, MANOVA, and ANCOVA techniques. *Review of Educational Research, 3,* 350–386.

Li, K. H., Raghunathan, T. E., & Rubin, D. B. (1991). Large-sample significance levels from multiple imputed data using moment-based statistics and an F distribution. *Journal of the American Statistical Association, 86,* 1065–1073.

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

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

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data.* Hoboken, NJ: Wiley.

Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhya, Ser. B, 36,* 115–128.

McQueen, A., Getz, J. G., & Bray, J. H. (2003). Acculturation, substance use, and deviant behavior: Examining separation and family conflict as mediators. *Child Development, 74,* 1737–1750.

Muthén, B., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52*(3), 431–462.

Muthén, L. K., & Muthén, B. O. (2004). *Mplus user's guide* (3rd ed.). Los Angeles: Muthén & Muthén.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Roth, P. L. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537–560.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* (63), 581–592.

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

Sadler, P., & Woody, E. (2003). Is who you are who you're talking to? Interpersonal style and complementarity in mixed-sex interactions. *Journal of Personality and Social Psychology, 84,* 80–96.

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

Schafer, J. L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model [Computer software]. University Park, PA: Department of Statistics, Pennsylvania State University.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177.

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.

Sinharay, S., Stern, H. S., & Russell, D. (2001). The use of multiple imputation for the analysis of missing data. *Psychological Methods, 6*(4), 317–329.

Snyder, J., Brooker, M., Patrick, M. R., Snyder, A., Schrepferman, L., & Stoolmiller, M. (2003). Observed peer victimization during early elementary school: Continuity, growth, and relation to risk for child antisocial and depressive behavior. *Child Development, 74,* 1881–1898.

Thompson, B. (1999, April). *Common methodology mistakes in educational research, revisited, along with a primer on both effect sizes and the bootstrap.* Invited address presented at the annual meeting of the American Educational Research Association, Montreal. (ERIC Document Reproduction Service No. ED 429 110)

Tolan, P. H., Gorman-Smith, D., & Henry, D. B. (2003). The developmental ecology of urban males' youth violence. *Developmental Psychology, 39,* 274–291.

West, S. G. (2001). New approaches to missing data in psychological research: Introduction to the special section. *Psychological Methods, 6,* 315–316.

Wilkinson, L., & Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist, 54,* 594–604.

Wothke, W. (2000). Longitudinal and multi-group modeling with missing data. In T. D. Little, K. U. Schnabel, & J. Baumert (Eds.), *Modeling longitudinal and multiple group data: Practical issues, applied approaches and specific examples* (pp. 219–240). Mahwah, NJ: Lawrence Erlbaum.

Yuan, K.-H., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. In M. Becker & M. Sobel (Eds.), *Sociological methodology, 2000* (pp. 165–200). Malden, MA: Blackwell.

Authors

JAMES L. PEUGH is a doctoral candidate in the Quantitative and Qualitative Methods in Education Program in the Educational Psychology Department, University of Nebraska, 114 Teachers College Hall, Lincoln, NE 68588-0345; e-mail *oaktreetx@yahoo.com.* His research interests include multilevel structural equation models, and he is director of the Nebraska Evaluation and Research Center at the University of Nebraska.

CRAIG K. ENDERS is an Assistant Professor in the Quantitative and Qualitative Methods in Education Program, in the Educational Psychology Department, University of Nebraska, 222 Teachers College Hall, Lincoln, NE 68588-0345; e-mail *cenders@unl.edu.* His research interests include missing data handling techniques, structural equation modeling, longitudinal modeling, and growth mixture models.