Missing Data Analysis: Making It Work in the Real World

John W. Graham

Department of Biobehavioral Health and the Prevention Research Center, The Pennsylvania State University, University Park, Pennsylvania 16802; email: jgraham@psu.edu

Annu. Rev. Psychol. 2009. 60:549–576

First published online as a Review in Advance on July 24, 2008

The *Annual Review of Psychology* is online at psych.annualreviews.org

This article's doi:
10.1146/annurev.psych.58.110405.085530

0066-4308/09/0110-0549$20.00

Key Words

multiple imputation, maximum likelihood, attrition, nonignorable missingness, planned missingness

Abstract

This review presents a practical summary of the missing data literature, including a sketch of missing data theory and descriptions of normal-model multiple imputation (MI) and maximum likelihood methods. Practical missing data analysis issues are discussed, most notably the inclusion of auxiliary variables for improving power and reducing bias. Solutions are given for missing data challenges such as handling longitudinal, categorical, and clustered data with normal-model MI; including interactions in the missing data model; and handling large numbers of variables. The discussion of attrition and nonignorable missingness emphasizes the need for longitudinal diagnostics and for reducing the uncertainty about the missing data mechanism under attrition. Strategies suggested for reducing attrition bias include using auxiliary variables, collecting follow-up data on a sample of those initially missing, and collecting data on intent to drop out. Suggestions are given for moving forward with research on missing data and attrition.

Contents

INTRODUCTION AND OVERVIEW

Missing data have challenged researchers since the beginnings of field research. The challenge has been particularly acute for longitudinal research, that is, research involving multiple waves of measurement on the same individuals. The main issue is that the analytic procedures researchers use, many of which were developed early in the twentieth century, were designed to have complete data. Until relatively recently, there was simply no mechanism for handling the responses that were sometimes missing within a particular survey, or whole surveys that were missing for some waves of a multiwave measurement project.

Problems brought about by missing data began to be addressed in an important way starting in 1987, although a few highly influential articles did appear before then (e.g., Dempster et al. 1977, Heckman 1979, Rubin 1976). What happened in 1987 was nothing short of a revolution in thinking about analysis of missing data. The revolution began with two major books that were published that year. Little & Rubin (1987) published their classic book, *Statistical Analysis with Missing Data* (the second edition was published in 2002). Also, Rubin

(1987) published his book, *Multiple Imputation for Nonresponse in Surveys*. These two books, coupled with the advent of powerful personal computing, would lay the groundwork for missing data software to be developed over the next 20 years and beyond. Also published in 1987 were two articles describing the first truly accessible method for dealing with missing data using existing structural equation modeling (SEM) software (Allison 1987, Muthén et al. 1987). Finally, Tanner & Wong (1987) published their article on data augmentation, which would become a cornerstone of the multiple imputation (MI) software that would be developed a decade later.

Goals of This Review

A major goal of this review is to present ideas and strategies that will make missing data analyses useful to researchers. My aim here is to encourage researchers to use the missing data procedures that are already known to be good ones. Efforts toward this goal involve summarizing the major research in missing data analysis over the past several years. However, much of the reluctance to adopt these procedures is related to the myths and misconceptions that continue to abound about the impact of missing data with and without using these procedures. Thus, a goal of this review is to clear up many of the myths and misconceptions surrounding missing data and analysis with missing data.

Work is required to become a practiced user of the acceptable (i.e., MI and maximum-likelihood, or ML) procedures. But that work would be a lot less onerous if one had confidence that learning these procedures would truly make one's work better and that criticisms surrounding missing data would be materially reduced.

Researchers should use MI and ML procedures (see Schafer & Graham 2002). They are good procedures that are based on strong statistical traditions. They can certainly be improved on, but by how much? I would argue that using MI and ML procedures gets us at least 90% of the way to the hypothetical ideal from where we were 25 years ago. Newer procedures will continually fine-tune the existing MI and ML procedures, but the main missing data solutions are already available and should be used now.

Above all, my goal is that this review will be of practical value. I hope that my words will facilitate the use of MI and ML missing data methods. This is not intended to be a thorough review of all work and methods relating to missing data. I have focused on what I believe to be most useful.

What's to Come

In the following sections, I discuss three major missing data topics: missing data theory, analysis in practice, and attrition and missingness that is not missing at random. In the first major section, I lay out the main tenets of what I refer to as "missing data theory." One central focus in this section is the causes or mechanisms of missingness. In this section, I discuss what I refer to as the "old" methods for dealing with missing data, but as much as possible, my discussion is limited to methods that remain useful at least in some circumstances. This section briefly presents the methods I fully endorse: MI and ML.

In the second major section, I focus on the practical side of performing missing data analyses. Over the years, I have faced all of these problems as a data analyst; these are real solutions. Sometimes the solutions are a bit ad hoc. Better solutions may become available in the future, but the solutions I present are known to have minimal harmful impact on statistical inference, and they will keep you doing analysis, which is the most important thing. In this section, I also touch on the developing area of planned missingness designs, an area that opens up new design possibilities for researchers who are already making use of the recommended MI and ML missing data procedures. Contrary to the old adage that the best solution to missing data is not to have them, there are times when building missing data into the overall measurement design is the best use of limited resources.

In the final major section, I describe the area of attrition and missingness that is not missing at random. This kind of missingness has proven to be a major obstacle, especially in longitudinal and intervention research. A good bit of the problem in this area stems from the fact that the framework for thinking about these issues was developed and solidified well before the missing data revolution. In this section, I propose a different framework for thinking about attrition and make several suggestions (pleas?) as to how researchers might proceed in this area.

MISSING DATA THEORY

Causes or Mechanisms of Missingness

Statisticians talk about missingness mechanisms. But what they mean by that term differs from what social and behavioral scientists think of as mechanisms. When I (trained as an experimental social psychologist) use that word, I think of causal mechanisms. What is the reason the data are missing? Statisticians, on the other hand, often are thinking more along the lines of a description of the missingness. For example, it is not uncommon to talk about a vector *R* for each variable, which takes on the value "1" if the variable has data for that case, and "0" if the value is missing for that case. This leads naturally to descriptions of the missing data, that is, patterns of missingness. For example, suppose that one has three variables (X, Y_1, and Y_2), and suppose that X is never missing but Y_1 is missing for some individuals, and Y_2 is missing for a few more. Or, thinking about it the other way, suppose one has data for all three variables for some number of cases, but partial data (X and Y_1) for some number of cases and partial data (X only) for some other number of cases. The patterns of missingness for a hypothetical N = 100 cases might look like those shown in **Table 1**. Also, as shown in **Table 1**, it not uncommon for a small number of cases to be present at one wave, missing at a later wave, and then give data at a still later wave.

Table 1 Hypothetical patterns of missingness

Variable			
X	Y_1	Y_2	N
1	1	1	65
1	1	0	20
1	0	0	10
1	0	1	5

1 = value present; 0 = value missing.

When people talk about the mechanisms of missingness, three terms come up: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). Although statisticians prefer not to use the word "cause," they do often use the words "due to" or "depends on" in this context.

With MAR, the missingness (i.e., whether the data are missing or not) may depend on observed data, but not on unobserved data (Schafer & Graham 202).

MCAR is a special case of MAR in which missingness does not depend on the observed data either (Schafer & Graham 2002).

With MNAR, missingness does depend on unobserved data.

More about MAR, MCAR, and MNAR. Although these three important terms do have specific statistical definitions, their practical meaning is often elusive. MCAR is perhaps the easiest to understand. If the cases for which the data are missing can be thought of as a random sample of all the cases, then the missingness is MCAR. This means that everything one might want to know about the data set as a whole can be estimated from any of the missing data patterns, including the pattern in which data exist for all variables, that is, for complete cases (e.g., see the top row in **Table 1**).

Another aspect of MCAR that is particularly easy to understand for psychologists is that the word "random" in MCAR means what psychologists generally think of when they use the term. The word "random" in MAR, however, means something rather different from what psychologists typically think of as random. In fact, the randomness in MAR missingness means that once one has conditioned on (e.g., controlled for) all the data one has, any remaining missingness is completely random (i.e., it does not depend on some unobserved variable). Because of this, I often say that a more precise term for missing at random would be conditionally missing at random. However, if such a term were in common use, its acronym (CMAR) would often be confused with MCAR. Thus, my feeling about this is that psychologists should continue to refer to MCAR and MAR but simply understand that the latter term refers to conditionally missing at random.

Another distinction that is often used with missingness is the distinction between ignorable and nonignorable missingness. Without going into detail here, suffice it to say that ignorable missingness applies to MCAR and MAR, whereas nonignorable missingness is often used synonymously with MNAR.

An important wrinkle with these terms, especially MAR and MNAR (or ignorable and nonignorable), is that they do not apply just to the data. Rather they apply jointly to the data and to the analysis that is being used. For example, suppose one develops a smoking prevention intervention and has a treatment group and a control group (represented by a dummy variable called Program). Suppose that one measures smoking status at time 2, one year after implementation of the prevention intervention (Smoking_2). Finally, suppose that some people have missing data for Smoking_2, and that missingness on Smoking_2 depends on smoking status measured at time 1, just before the program implementation (Smoking_1). If one includes Smoking_1 in one of the acceptable missing data procedures (MI or ML), then the missingness on Smoking_2 is conditioned on Smoking_1 and is thus MAR (note that even with complete case analysis, the regression analysis of Program predicting Smoking_2 is MAR as long as Smoking_1 is also included in the model; e.g., see Graham & Donaldson 1993). However, if the researcher tested a model in which the Program alone predicted Smoking_2, then the missingness would become MNAR because the researcher failed to condition on Smoking_1, the cause of missingness.

Many researchers have suggested modifying the names of the missing data mechanisms in order to have labels that are a bit closer to regular language usage. However, missing data theorists believe that these mechanism names should remain as is. I agree. I now believe we would be doing psychologists a disservice if we encouraged them to abandon these terms, which are so well entrenched in the statistics literature. Rather, we should continue to use these terms (MCAR, MAR, and MNAR), but always define them very carefully using regular language.

Consequences of MCAR, MAR, and MNAR. The main consequence of MCAR missingness is loss of statistical power. The good thing about MCAR is that analyses yield unbiased parameter estimates (i.e., estimates that are close to population values). MAR missingness (i.e., when the cause of missingness is taken into account) also yields unbiased parameter estimates. The reason MNAR missingness is considered a problem is that it yields biased parameter estimates (discussed at length below).

Old Analyses: A Brief Summary

This summary is not intended to be a thorough examination of the "old" approaches for dealing with missing data. Rather, in order to be of most practical value, the discussion below focuses on the old approaches that can still be useful, at least under some circumstances.

Yardsticks for evaluating methods. I have judged the various methods (old and new) by three means. First, the method should yield unbiased parameter estimates over a wide range of parameters. That is, the parameter estimate should be close to the population value for that parameter. Some of the methods I would judge to be unacceptable (e.g., mean substitution) may yield a mean for a particular variable that is close to the true parameter value (e.g., under MCAR), but other parameters using this method can be seriously biased. Second, there should be a method for assessing the degree of uncertainty about parameter estimates. That is, one should be able to obtain reasonable estimates of the standard error or confidence intervals. Third, once bias and standard errors have been dealt with, the method should have good statistical power.

Complete cases analysis (AKA, listwise deletion). This approach can be very useful even today. One concern about listwise deletion is that it may yield biased parameter estimates (e.g., see Wothke 2000). For example, groups with complete data, especially in a longitudinal study, often are quite different from those that have missing data. Nevertheless, the difference in those two groups often is embodied completely in the pretest variables (for which everyone has data). Thus, as long as those variables can reasonably be included in the model as covariates, the bias is often minimal, even with listwise deletion, especially for multiple regression models (e.g., see Graham & Donaldson 1993).

However, there will always be some loss of power with listwise deletion because of the unused partial data. And in some instances, this loss of power can be huge, making this method an undesirable option. Still, if the loss of cases due to missing data is small (e.g., less than about 5%), biases and loss of power are both likely to be inconsequential. I, personally, would still use one of the missing data approaches even with just 5% missing cases, and I encourage you to get used to doing the same. However, if a researcher chose to stay with listwise deletion under these special circumstances, I believe it would be unreasonable for a critic to argue that it was a bad idea to do so. It is also important that standard errors based on listwise deletion are meaningful.

POSITIVE DEFINITE

One good way to think of a matrix that is not positive definite is that the matrix contains less information than implied by the number of variables in the matrix. For example, if a matrix contained the correlations of three variables (*A*, *B*, and *C*) and their sum, then there would be just three variables worth of information, even though it contained four variables. Because the sum is perfectly predicted by the three variables, it adds no new information, and the matrix would not be positive definite.

Pairwise deletion. Pairwise deletion is usually used in conjunction with a correlation matrix. Each correlation is estimated based on the cases having data for both variables. The issue with pairwise deletion is that different correlations (and variance estimates) are based on different subsets of cases. Because of this, it is possible that parameter estimates based on pairwise deletion will be biased. However, in my experience, these biases tend to be small in empirical data. On the other hand, because different correlations are based on different subsets of cases, there is no guarantee that the matrix will be positive definite (see sidebar Positive Definite). Nonpositive definite matrices cannot be used for most multivariate statistical analyses. A bigger concern with pairwise deletion is that there is no basis for estimating standard errors.

Because of all these problems, I cannot recommend pairwise deletion as a general solution. However, I do still use pairwise deletion in one specific instance. When I am conducting preliminary exploratory factor analysis with a large number of variables, and publication of the factor analysis results, per se, is not my goal, I sometimes find it useful to conduct this analysis with pairwise deletion. As a preliminary analysis for conducting missing data analysis, I sometimes examine the preliminary eigenvalues from principal components analysis (see sidebar Eigenvalue). If the last eigenvalue is positive, then the matrix is positive definite. Many failures of the expectation-maximization (better known as EM) algorithm (and MI) are due to the correlation matrix not being positive definite.

Other "old" methods. Other old methods include mean substitution, which I do not recommend. In the "Modern" Missing Data Analysis

Methods section, I describe a method based on the EM algorithm that is much preferred over mean substitution (see the sections on Good Uses of the EM Algorithm and Imputing a Single Data Set from EM Parameters). A second method that has been described in the literature involves using a missingness dummy variable in addition to the specially coded missing value. This approach has been discredited and should not be used (e.g., see Allison 2002). Finally, regression-based single imputation has been employed in the past. Although the concept is a sound one and is the basis for many of the modern procedures, this method is not recommended in general.

"Modern" Missing Data Analysis Methods

The "modern" missing data procedures I consider here are (*a*) the EM algorithm, (*b*) multiple imputation under the normal model, and (*c*) ML methods, often referred to as full-information maximum likelihood (FIML) methods.

EM algorithm. It is actually a misnomer to refer to this as "the" EM algorithm because there are different EM algorithms for different applications. Each version of the EM algorithm reads in the raw data and reads out a different product, depending on the application. I describe here the EM algorithm that reads in the raw data, with missing values, and reads out an ML variance-covariance matrix and vector of means. Definitive technical treatments of various EM algorithms are given in Little & Rubin (1987, 2002) and Schafer (1997). Graham and colleagues provide less-technical descriptions of the workings of the EM algorithm for covariance matrices (Graham & Donaldson 1993; Graham et al. 1994, 1996, 1997, 2003).

In brief, the EM algorithm is an iterative procedure that produces maximum likelihood estimates. For the E-step at one iteration, cases are read in, one by one. If a value is present, the sums, sums of squares, and sums of cross-products are incremented. If the value is missing, the current best guess for that value is used instead. The best guess is based on regression-based single imputation with all other variables in the model used as predictors. For the sums, the best guess value is used as is. For sums of squares and sums of cross-products, if just one value is missing, then the quantity is incremented directly. However, if both values are missing, then the quantity is incremented, and a correction factor is added. This correction is conceptually equivalent to adding a random residual error term in MI (described below).

In the M-step of the same iteration, the parameters (variances, covariances, and means) are estimated (calculated) based on the current values of the sums, sums of squares, and sums of cross-products. Based on the covariance matrix at this iteration, new regression equations are calculated for each variable predicted by all others. These regression equations are then used to update the best guess for missing values during the E-step of the next iteration. This two-step process continues until the elements of the covariance matrix stop changing. When the changes from iteration to iteration are so small that they are judged to be trivial, EM is said to have converged.

Being ML, the parameter estimates (means, variances, and covariances) from the EM algorithm are excellent. However, the EM algorithm does not provide standard errors as an automatic part of the process. One could obtain an estimate of these standard errors using bootstrap procedures (e.g., see Graham et al. 1997). Although bootstrap procedures (Efron

EIGENVALUE

Eigenvalues are part of the decomposition of a correlation matrix during factor analysis or principal components analysis. Each eigenvalue represents the variance of the linear combination of items making up that factor. In principal components, the total variance for the correlation matrix is the number of items in the matrix. If the matrix is positive definite, the last eigenvalue will be positive, that is, it will have variance. However, if the matrix is not positive definite, one or more of the eigenvalues will be 0, implying that those factors have no variance, or that they add no new information over and above the other factors.

1982) are often criticized, they can be quite useful in this context as a means of dealing with nonnormal data distributions.

Good uses of the EM algorithm. Although the EM algorithm provides excellent parameter estimates, the lack of convenient standard errors means that EM is not particularly good for hypothesis testing. On the other hand, several important analyses, often preliminary analyses, don't use standard errors anyway, so the EM estimates are very useful. First, it is often desirable to report means, standard deviations, and sometimes a correlation matrix in one's paper. I would argue that the best estimates for these quantities are the ML estimates provided by EM. Second, data quality analyses, for example, coefficient alpha analyses, because they typically do not involve standard errors, can easily be based on the EM covariance matrix (e.g., see Enders 2003; Graham et al. 2002, 2003). The EM covariance matrix is also an excellent basis for exploratory factor analysis with missing data. This is especially easy with the SAS/STAT® software program (SAS Institute); one simply includes the relevant variables in Proc MI, asking for the EM matrix to be output. That matrix may then be used as input for Proc Factor using the "type = cov" option.

Although direct analysis of the EM covariance matrix can be useful, a more widely useful EM tool is to impute a single data set from EM parameters (with random error). This procedure has been described in detail in Graham et al. (2003). This single imputed data set is known to yield good parameter estimates, close to the population average. But more importantly, because it is a complete data set, it may be read in using virtually any software, including SPSS. Once read into the software, coefficient alpha and exploratory factor analyses may be carried out in the usual way. One caution is that this data set should not be used for hypothesis testing. Standard errors based on this data set, say from a multiple regression analysis, will be too small, sometimes to a substantial extent. Hypothesis testing should be carried out with MI or one of the FIML procedures. Note that the procedure in SPSS for writing out a single imputed data set based on the EM algorithm is not recommended unless random error residuals are added after the fact to each imputed value; the current implementation of SPSS, up to version 16 at least, writes data out without adding error (e.g., see von Hippel 2004). This is known to produce important biases in the data set (Graham et al. 1996).

Implementations of the EM algorithm. Good implementations of the EM algorithm for covariance matrices are widely available. SAS Proc MI estimates the EM covariance matrix as a by-product of its MI analysis (SAS Institute). Schafer's (1997) NORM program, a stand-alone Microsoft Windows program, also estimates the EM covariance matrix as a step in the MI process. Graham et al. (2003) have described utilities for making use of that covariance matrix. Graham & Hofer (1992) have created a stand-alone DOS-based EM algorithm, EMCOV, which can be useful in simulations.

Multiple imputation under the normal model. I describe in this section MI under the normal model as it is implemented in Schafer's (1997) NORM program. MI as implemented in SAS Proc MI is also based on Schafer's (1997) algorithms, and thus is the same kind of program as NORM. Detailed, step-by-step instructions for running NORM are available in Graham et al. (2003; also see Graham & Hofer 2000, Schafer 1999, Schafer & Olsen 1998). Note that Schafer's NORM program is also available as part of the Splus missing data library (**http://www.insightful.com/**).

The key to any MI program is to restore the error variance lost from regression-based single imputation. Imputed values from single imputation always lie right on the regression line. But real data always deviate from the regression line by some amount. In order to restore this lost variance, the first part of imputation is to add random error variance (random normal error in this case). The second part of restoring lost variance relates to the fact that each imputed value is based on a single regression equation, because the regression equation, and the underlying covariance matrix, is based on a single draw from the population of interest.

In order to adjust the lost error completely, one should obtain multiple random draws from the population and impute multiple times, each with a different random draw from the population. Of course, this is almost never possible; researchers have just a single sample. One option might be to simulate random draws from the population by using bootstrap procedures (Efron 1982). Another approach is to simulate random draws from the population using data augmentation (DA; Tanner & Wong 1987).

The key to Schafer's NORM program is DA. NORM first runs EM to obtain starting values for DA. DA can be thought of as a kind of stochastic (probabilistic) version of EM. It, too, is an iterative, two-step process. There is an imputation step during which DA simulates the missing data based on the current parameter estimates, and a posterior step during which DA simulates the parameters given the current (imputed) data. DA is a member of the Markov-Chain Monte Carlo family of algorithms. It is Markov-like in the sense that all of the information from one step of DA is contained in the previous step. Because of this, the parameter estimates and imputed values from two adjacent steps of DA are more similar than one would expect from two random draws from the population. However, after, say, 50 steps of DA, the parameter estimates and imputed values from the initial step and those 50 steps removed are much more like two random draws from the population. The trick is to determine how many steps are required before the two imputed data sets are sufficiently similar to two random draws from the population. Detailed guidance for this process is given in Graham et al. (2003). In general, the number of iterations it takes EM to converge is an excellent estimate of the number of steps there should be between imputed data sets from DA (this rule applies best to the NORM program; different MI programs have different convergence criteria, and this rule may be slightly different with those MI programs). In addition, diagnostics are available in NORM and Proc MI to verify that the number of DA steps selected was good enough.

Implementations of MI under the normal model. Implementations of MI under the normal model are also widely available. Schafer's (1997) NORM software is a free program (see **http://methodology.psu.edu/** for the free download). SAS Proc MI (especially version 9, but to a large extent version 8.2; SAS Institute) provides essentially the same features as NORM. For analyses conducted in SAS, Proc MI is best. Other implementations of MI are not guaranteed to be as robust as are those based on DA or other Markov-Chain Monte Carlo routines, although such programs may be useful under specific circumstances. For example, Amelia II (see Honaker et al. 2007, King et al. 2001) and IVEware (Raghunathan 2004) are two MI programs that merit a look. See Horton & Kleinman (2007) for a recent review of MI software.

Special MI software for categorical, longitudinal/cluster, and semi-continuous data. In this category are Schafer's (1997) CAT program (for categorical data) and MIX program (for mixed continuous and categorical problems). Both of these are available (along with NORM) as special commands in the latest version of Splus. Although CAT can certainly be used to handle imputation with categorical data, it presents the user with some limitations. Most importantly, the default in CAT involves what amounts to the main effects and all possible interactions. Thus, even with a few variables, the default model can involve a huge number of parameters. For example, with just five input variables, CAT estimates parameters for 31 variables (five main effects, ten 2-way interactions, ten 3-way interactions, five 4-way interactions, and one 5-way interaction).

Also included in this category is the PAN program (for special panel and cluster-data designs, see Schafer 2001, Schafer & Yucel 2002). PAN was created for the situation in which a variable, Posatt (beliefs about the positive social consequences of alcohol use), was measured in grades 5, 6, 7, 9, and 10 of a longitudinal study, but was omitted for all subjects in grade 8. Other variables (e.g., alcohol use), however, were measured in all six grades. Because no subject had data for Posatt measured in eighth grade, MI under the normal model could not be used to impute that variable. However, because PAN also takes into account growth (change) over time, Posatt 8 can be imputed with PAN. PAN is also very good for imputing clustered data (e.g., students within schools) where the number of clusters is large. Although the potential for the PAN program is huge, its availability remains limited.

Olsen & Schafer (2001) have described multiple imputation for semi-continuous data, especially in the growth modeling context. Semi-continuous data come from variables that have many responses at one value (e.g., 0) and are more or less normally distributed for values greater than 0.

Imputing a single data set from EM parameters. An often useful alternative to analyzing the EM covariance matrix directly is to impute a single data set based on EM parameters (+random error, an option available in Schafer's NORM program; see Graham et al. 2003 for details). With data sets imputed using data augmentation, parameter estimates can be anywhere in legitimate parameter space. However, when the single imputation (+error) is based on the EM covariance matrix, all parameter estimates are near the center of the parameter space. For this reason, if one analyzes just one imputed data set, it should be this one. This data set is very useful for analyses that do not require hypothesis testing, such as coefficient alpha analysis and exploratory factor analysis (see Graham et al. 2003 for additional details).

FIML methods. FIML methods deal with the missing data, do parameter estimation, and estimate standard errors all in a single step. This means that the regular, complete-cases algorithms must be completely rewritten to handle missing data. Because this task is somewhat daunting, software written with the FIML missing data feature is limited. At present, the feature is most common in SEM software (in alphabetical order, Amos: Arbuckle & Wothke 1999; LISREL: Jöreskog & Sörbom 1996, also see du Toit & du Toit 2001; Mplus: Muthén, & Muthén 2007; and Mx: Neale et al. 1999). Although each of these programs was written specifically for SEM applications, they can be used for virtually any analysis that falls within the general linear model, most notably multiple regression. For a review of FIML SEM methods, see Enders (2001a).

Other FIML (or largely FIML) software for latent class analysis includes Proc LTA (e.g., Lanza et al. 2005; also see **http://methodology.psu.edu**) and Mplus (Muthén & Muthén 2007).

Other "older" methods. One other method deserves special mention in this context. Although SEM analysis with missing data is currently handled almost exclusively by SEM/FIML methods (see previous section), an older method involving the multiple group capabilities of SEM programs is very useful for some applications. This approach was described initially by Allison (1987) and Muthén et al. (1987). Among other things, this method has proven to be extremely useful with simulations involving missing data (e.g., see Graham et al. 2001, 2006).

This method also continues to be useful for measurement designs described as "accelerated longitudinal" or "cohort sequential" (see Duncan & Duncan 1994; Duncan et al. 1994, 1996; McArdle 1994; McArdle & Hamagami 1991, 1992). With these designs, one collects data for two or more sets of participants of different ages over, say, three consecutive years. For example, one group is 10, 11, and 12 years old over the three years of a study, and another group is 11, 12, and 13 years old over the same three study years. Because no participants have data for both ages 10 and 13, regular FIML-based SEM software and normal-model MI cannot be used in this context. However, the multiple-group SEM approach may be used to test a growth model covering growth over all four ages.

Dispelling Myths About MAR Missing Data Methods

Myths abound regarding missing data and analysis with missing data. Many of these myths originated with thinking that was developed well before the missing data revolution. Parts of that earlier thinking, of course, remain an important element of modern psychological science. But the parts relating to missing data need to be revised. I address three of the most common myths in this section. Other myths are dealt with in the sections that follow.

Imputation is making up the data. It is true that imputation is the process of plugging in plausible values where none exist. But the point of this process is not to obtain the individual values themselves. Rather, the point is to plug in these values (multiple times) in order to preserve important characteristics of the data set as a whole. By "preserve," I mean that parameter estimates should be unbiased. That is, the estimated mean, for example, should be close to the true population value for the mean; the estimated variance should be close to the true population value for the variance. In this review, I talk mainly about multiple imputation under the normal model. Normal-model MI "preserves" means, variances, covariances, correlations, and linear regression coefficients.

You are unfairly helping yourself by imputing (AKA, it is okay to impute the independent variable, but not the dependent variable). There are several versions of this myth. In the past, some researchers were convinced that imputation procedures such as normal-model MI were fine for imputing missing data that might occur within the set of independent variables (IVs) (and covariates) of a study. However, these researchers were very reluctant to include the dependent variable (DV) in the MI model when it, too, included missing values. They felt that it was somehow unfair to impute the DV.

The truth is that all variables in the analysis model must be included in the imputation model. The fear is that including the DV in the imputation model might lead to bias in estimating the important relationships (e.g., the regression coefficient of a program variable predicting the DV). However, the opposite actually happens. When the DV is included in the model, all relevant parameter estimates are unbiased, but excluding the DV from the imputation model for the IVs and covariates can be shown to produce biased estimates. The problem with leaving the DV out of the imputation model is this: When any variable is omitted from the model, imputation is carried out under the assumption that the correlation is $r = 0$ between the omitted variable and variables included in the imputation model. Thus, when the DV is omitted, the correlations between it and the IVs (and covariates) included in the model are all suppressed (i.e., biased) toward 0.

MAR methods don't work if the MAR assumption does not hold (AKA, complete cases are preferred if MAR does not hold). With some procedures, such as multiple linear regression, it is assumed that the data are multivariate normal. Violation of this assumption is known to affect the results (most notably the standard errors of the regression coefficients). So with multiple regression, if the normality assumption has been violated, one should use a different procedure. This logic makes good sense with multiple regression analysis, but it does not apply to analysis with missing data because with multiple regression, when the normality assumption is violated, other common procedures work better. But with missing data, when the MAR assumption has been violated, the violation affects the old procedures (e.g., listwise deletion) as well, and typically this violation has greater effect on the old procedures. In short, MI and ML methods are always at least as good as the old procedures (e.g., listwise deletion, except in artificial, unrealistic circumstances), and MI/ML methods are typically better than old methods, and often very much better.

An important difference between MI/ML methods and complete cases analysis is that auxiliary variables (see next section) may be used with MI/ML in order to reduce the impact of MNAR missingness. However, there is no good way of incorporating auxiliary variables into a complete cases model unless they can reasonably be incorporated (e.g., as covariates) into the model of substantive interest.

PRACTICAL ISSUES: MAKING MISSING DATA ANALYSIS WORK IN THE REAL WORLD

The suggestions given here are designed to make missing data analyses useful in real-world data situations. Some of the suggestions given here are necessarily brief. Many other practical suggestions are given in Graham et al. (2003) and elsewhere.

Inclusive versus Restrictive Variable Inclusion Strategies (MI versus FIML)

In some ways, this is the most important lesson that can be learned when doing missing data analysis in the real world. Collins et al. (2001) discussed the differences between "inclusive" and "restrictive" variable inclusion strategies in missing data analysis. An inclusive strategy is one in which auxiliary variables are included in the model. An auxiliary variable is a variable that is not part of the model of substantive interest, but is highly correlated with the variables in the substantive model. Collins et al. (2001) showed that including auxiliary variables in the missing data model can be very helpful in two important ways. It can reduce estimation bias due to MNAR missingness, and it can partially restore lost power due to missingness.

Collins et al. (2001) note that the potential auxiliary variable benefit is the same for MI and FIML analyses but that the typical use of MI is different from typical use of FIML. For MI analyses, including auxiliary variables in the imputation model has long been practiced and is very easy to accomplish: Simply add the variables to the imputation model. Furthermore, once the auxiliary variables have been included in the imputation model, subsequent analyses involving the imputed data benefit from the auxiliary variables, whether or not those variables appear in the analysis of substantive interest (this latter benefit also applies to analysis of the EM covariance matrix). On the other hand, FIML analyses have typically included only the variables that are part of the model of substantive interest. Thus, researchers who use FIML models have found it difficult to incorporate auxiliary variables in a reasonable way. Fortunately, reasonable approaches for including auxiliary variables into SEM/FIML models now exist (e.g., see Graham 2003; also see the recently introduced feature in Mplus for easing the process of including auxiliary variables). Note that although these methods work well for SEM/FIML models, no corresponding strategies are available at present for incorporating auxiliary variables into latent class FIML models.

Small Sample Sizes

Graham & Schafer (1999) showed that MI performs very well in small samples (as low as N = 50), even with very large multiple regression models (as large as 18 predictors) and even with as much as 50% missing data in the DV. The biggest issue with such small samples is not the missingness, per se, but rather that one simply does not have much data to begin with and missingness depletes one's data even further. MI was shown to perform very well under these circumstances; the analyses based on MI data were as good as the same analyses performed on complete data.

The simulations performed by Graham & Schafer (1999) also showed that normal-model MI with nonnormal data works well, as well as analysis with the same data with no missing values. Although analysis of imputed data works as well as analysis with complete datasets, nothing in the imputation process, per se, fixes the nonnormal data. Thus, in order to correct the problems with standard errors often found with nonnormal data, analysis procedures must be used that give correct standard errors (e.g., the correction given for SEM by Satorra & Bentler

1994). Enders (2001b) has drawn similar conclusions regarding the use of the FIML missing data feature for SEM programs.

Rounding

Rounding should be kept to a minimum. MI was designed to restore the lost variability found in single imputation, and the MI strategy was designed to yield the correct variability. Rounding is tantamount to adding more variability to the imputed values. The added variability is random, to be sure, but there is definitely more of it with rounding than without. This additional variance is evident in coefficient alpha analyses with rounded and unrounded imputed values in a single data set imputed from EM parameters. Coefficient alpha is always a point or two lower (showing more random error variance) with rounding than without (also, see the Categorical Missing Data and Normal-Model MI section below for a discussion regarding rounding for categorical variables).

Number of Imputations in MI

Missing data theorists have often claimed that good inferences can be made with the number of imputed data sets (*m*) as few as *m* = 3 to 5. They have argued that the relative efficiency of estimation is very high under these circumstances, compared to an infinite number of imputations. However, Graham et al. (2007) have recently shown that the effects of *m* on statistical power for detecting a small effect size ($\rho = 0.10$) can be strikingly different from what is observed for relative efficiency. They showed that if statistical power is the main consideration, the number of imputations typically must be much higher than previously thought. For example, with 50% missing information, Graham et al. (2007) showed that MI with *m* = 5 has a 13% power falloff compared to the equivalent FIML analysis; with 30% missing information and *m* = 5, there was a 7% power falloff compared to FIML. Graham et al. (2007) recommend that at least *m* = 40 imputations are needed with 50% missing information to guarantee less than a 1% power falloff compared to the comparable FIML analysis.

Making EM (and MI) Perform Better (i.e., Faster)

Factors that affect the speed of MI are the same as those that affect the speed of EM, so I focus here on the latter. EM involves matrix manipulations to a large extent, so sample size has relatively little effect. However, the number of variables (*k*) affects EM tremendously. Consider that EM estimates [$k\,(k{+}1)/2 + k$] parameters [*k* variances, (k(k−1)/2) covariances, and *k* means]. That means that with 25, 50, 100, 150, and 200 variables, EM must estimate 350, 1325, 5150, 11,475, and 20,300 parameters, respectively. Note that as the number of variables gets large, the number of estimated parameters in EM gets huge. Although there is leeway here, I generally try to keep the total number of variables under 100 even with large sample sizes of N = 1000 or more. With smaller sample sizes, a smaller number of variables should be used.

Also affecting the speed of EM and MI is the amount of missing information (similar to, but not the same as, the amount of missing data). More missing information means EM converges more slowly. Finally, the distributions of the variables can affect speed of convergence. With highly skewed data, EM generally converges much more slowly. For this reason, it is often a good idea to transform the data (e.g., with a log transformation) prior to imputation. The imputed values can be back-transformed (e.g., using the antilog) after imputation, if necessary.

If EM is very slow to converge, for example, if it takes more than about 200 iterations, the speed of convergence can generally be improved. If EM converges in 200 iterations, then one should ask for 200 steps of data augmentation between each imputed data set. With *m* = 40 imputed data sets, one would need to run 40 × 200 = 8000 steps of DA. If EM converged in 1000 iterations, one would need to run 40 × 1000 = 40,000 steps of DA. The additional time can be substantial, especially if the time between iterations is large.

Including Interactions in the Missing Data Model

An issue that comes up frequently in missing data analysis has to do with omitting certain variables from the missing data model. This issue is sometimes referred to as being sure that the imputation model is at least as general as the analysis model. A clear example is a test of the effect of an interaction (e.g., the product) of two variables on some third variable. Because the product is a nonlinear combination of the two variables, it is not part of the regular linear imputation model. The problem with excluding such variables from the imputation model is that all imputation is done under the assumption that the correlation is $r = 0$ between the omitted variable and all other variables in the imputation model. In this case, the correlation between the interaction term and the DVs of interest will be suppressed toward 0. The solution is to anticipate any interaction terms and include the relevant product terms in the imputation model.

Another way to conceive of an interaction is to think of one of the variables as a grouping variable (e.g., gender). The interaction in this case means that the correlation between two variables is different for males and females. In the typical imputation model, one imputes under the model that all correlations are the same for females and males. A good way to impute under a model that allows these correlations to be different is to impute separately for males and females. The advantage of this approach is that all interactions involving gender can be tested during analysis even if a specific interaction was not anticipated beforehand. This approach works very well in program effects analyses. If the program and control groups are imputed separately, then it is possible to test any interaction involving the program dummy variable after the fact. One drawback to imputing separately within groups is that it cuts the sample size at least in half. This may be acceptable with a large sample. But if the sample is too small for this strategy, then including a few carefully selected product terms may be the best option.

Longitudinal Data and Special Longitudinal Missing Data Models

Missing data models have been created for handling special longitudinal data sets (e.g., the PAN program; Schafer 2001). Some people believe that programs such as PAN must be used to impute longitudinal data, for example, in connection with growth curve modeling (see "Modern" Missing Data Analysis Methods section above). However, this is not the case. It is easiest to see this by examining the various ways in which growth curve analyses can be performed. Special hierarchical linear modeling programs (e.g., HLM; Raudenbush & Bryk 2002) can be used for this purpose. However, standard SEM programs can also be used (e.g., see Willett & Sayer 1994). When analyses are conducted with these models under identical conditions (e.g., assuming homogeneity of error variances over time), the results of these two procedures are identical.

The key for the present review is that a variance-covariance matrix and vector of means provide all that is needed for performing growth modeling in SEM. Thus, any missing data procedure that preserves (i.e., estimates without bias) variances, covariances, and means is acceptable. This is exactly what results from the EM algorithm, and, asymptotically, with normal-model MI. In summary, MI under the normal model, or essentially equivalent SEM models with a FIML missing data feature, may safely be used in conjunction with longitudinal data.

Categorical Missing Data and Normal-Model MI

Although some researchers believe that missing categorical data requires special missing data procedures for categorical data, this is not true in general. The proportion of people giving the

"1" response for a two-level categorical variable coded "1" and "0" is the same as the mean for that variable. Thus, the important characteristics of this variable are preserved, even using normal-model MI. If the binary variable (e.g., gender) is to be used as a covariate in a regression analysis, then the imputed values should be used, as is, without rounding (see Rounding section above). If the binary variable must be used in analysis as a binary variable, then each imputed value should be rounded to the nearest observed value (0 or 1). There are variations on this rounding procedure (e.g., see Bernaards et al. 2007), but the simple rounding is known to perform very well in empirical data.

With normal-model MI (this also applies to SEM analysis with FIML), categorical variables with two levels may be used directly. However, categorical variables with more than two levels must first be dummy coded. If there are p levels in the categorical variable, then $p - 1$ dummy variables must be created to represent the categorical variable. For example, with a categorical variable with four levels, this dummy coding is completed as shown in **Table 2**.

If the original categorical variable has no missing data, then creating these dummy variables and using them in the missing data analysis is all that must be done. However, if the original categorical variable does have missing data, then imputation under the normal model may not work perfectly, and an ad hoc fix must be used. The problem is that dummy coding has precise meaning when all dummy-code values for a particular person are 0 or if there is exactly one 1 for the person. If there is a missing value for the original categorical variable, then all of the dummy variables will also be missing and it is possible that a missing value for two (or more) of the dummy variables could be imputed as a 1 after rounding. If any people have 1 values for more than one of these dummy variables, then the meaning of the dummy variables is changed.

Table 2 Example of dummy coding with four-level categorical variable

	Dummy variable		
Original category	D1	D2	D3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

If the number of "illegal" imputed values in this situation is small compared with the overall sample size, then one could simply leave them. However, a clever, ad hoc fix for the problem has been suggested by Paul Allison (2002). To employ Allison's fix, it is important to impute without rounding. Whenever there is an illegal pattern of imputed values (more than a single 1 for the dummy variables), the value 1 is assigned to the dummy variable with the highest imputed value, and 0 is assigned to all others in the dummy variable set. The results of this fix will be excellent under most circumstances.

Estimating proportions and frequencies with normal-model MI. Although normal-model MI does a good job of preserving many important characteristics of the data set as a whole, it is important to note that it does not preserve proportions and frequencies, except in the special case of a variable with just two levels (e.g., yes and no coded as 1 and 0), in which case the proportion of people giving the 1 response is the same as the mean and is thus preserved. However, consider the question, "How many cigarettes did you smoke yesterday?" (0 = none, 1 = 1–5, 2 = 6 or more). Researchers may be interested in knowing the proportion of people who have smoked cigarettes. This is the same as the proportion of people who did not respond 0, or one minus the proportion who gave the 0 response. Although the mean of this three-level smoking variable will be correct with normal-model MI, the proportion of people with the 0 response is not guaranteed to be correct unless the three-level smoking variable happens to be normally distributed (which is unlikely in most populations). This problem can be corrected simply by performing a separate EM analysis with the two-level version of this smoking variable (e.g., 0 versus other). The EM mean provides the correct proportion. Correct frequencies for all response categories, if needed, may be obtained in this case by recasting the three-level categorical variable as two dummy variables.

Normal-Model MI with Clustered Data

The term "clustered data" refers to the situation in which cases are clustered in naturally occurring groups, for example, students within schools. In this situation, the members of each cluster are often more similar to one another in some important ways than they are to members of other clusters. This type of multilevel structure requires special methods of analysis (e.g., Raudenbush & Bryk 2002; also see Murray 1998). These multilevel analysis models allow variable means to be different across the different clusters (random intercepts model), and sometimes they allow the covariances to be different in the different clusters (random slopes and intercepts model). If the analysis of choice is a random slopes and intercepts model, then, just as described above, the imputation model should involve imputing separately within each cluster. However, if the analysis of choice involves only random intercepts, then a somewhat easier option is available. In this situation, the cluster membership variable can be dummy coded. That is, for p clusters, $p - 1$ dummy variables would be specified (see **Table 2** for a simple dummy variable example).

With the dummy coding strategy, the $p - 1$ dummy variables are included in the imputation model. As long as the number of clusters is relatively small compared with the sample size, this dummy coding strategy works well. I have seen this strategy work well with as many as 35 dummy variables, although a smaller number is desirable. Remember that the number of dummy variables takes away from the number of substantive variables that reasonably can be used in the imputation model.

When the number of clusters is too high to work with normal-model MI, several options are available. A specialty MI model (such as PAN; Schafer 2001) can be employed, but that strategy can sometimes be costly in terms of learning new procedures. In addition, the performance of PAN has not been adequately evaluated at this time. Alternatively, the number of clusters can be reduced in a reasonable way. For example, if it is known that certain clusters have similar means, then these clusters could be combined (i.e., they would have the same dummy variable) prior to imputation (note that this kind of combining of clusters must be done within experimental groups). For every combination of this sort that can be reasonably formed, the number of dummy variables is reduced by one. I have even seen the strategy of performing a k-means cluster analysis on the key study variables using school averages as input. This type of analysis would help identify the clusters of clusters for which means on key variables are similar.

One factor to take into consideration when employing the dummy variable approach to handling cluster data is that sometimes variables, especially binary variables, that have very low counts (e.g., marijuana use among fifth graders) will be constants within one or more of the clusters. The data should be examined for this kind of problem prior to attempting the dummy variable approach. A good way to start is to perform a principal components analysis on the variables to be included in the imputation model, along with the dummy variables. If the last eigenvalue from this analysis is positive, the dummy coding strategy will most likely work.

Large Numbers of Variables

This problem represents perhaps the biggest challenge for missing data analyses in large field studies, especially longitudinal field studies. Consider that most constructs are measured with multiple items. Five constructs with four items per construct translates into 20 individual variables. Five waves of measurement produce 100 variables. If more constructs are included in the analysis, it is difficult to keep the total down to the $k = 100$ that I have recommended. It is even more difficult to keep the variables in the imputation model to a reasonable number if one has cluster data and is employing the dummy variable strategy (see previous section). Below, I describe briefly two strategies that have worked well in practice.

Imputing whole scales. If the analysis involves latent variable analysis (e.g., SEM) such that the individual variables must be part of the analysis, then imputing whole scales is not possible. However, analyses often involve the whole scales anyway, so imputing at that level is an excellent compromise. As long as study participants have data either for all or for none of the scale items, then this strategy is easy. The problem with using this strategy is how to deal with the partial data on the scale. Schafer & Graham (2002) suggested that forming a scale score based on partial data can cause problems in some situations, but may be fine in others. In my experience, forming a scale score based on partial data will be acceptable (*a*) if a relatively high proportion of variables are used to form the scale score (and never fewer than half of the variables), and (*b*) when the variables are consistent with the domain sampling model (Nunnally 1967), and (*c*) when the variables have relatively high coefficient alpha. Variables can be considered consistent with the domain sampling model when the item-total correlations (or factor loadings) for the variables within the scale are all similar. Conceptually, it must be reasonable that dropping one variable from the scale has essentially the same meaning as dropping any other variables from the scale. Otherwise, one must either discard any partial data or impute at the individual item level for that scale. Imputing at the scale level for even some of the study scales will often help with the problem of having too many variables.

Think FIML. Another strategy that works well in practice is to start with the variables that would be included in the analysis model of substantive interest. If the FIML approaches are used, then these are the only variables that would be included using the typical practice. In addition, one should carefully select the few auxiliary variables that will have the most beneficial impact on the model. Adding auxiliary variables that are correlated $r = 0.50$ or better with the variables of interest will generally help the analysis to have less bias and more power. However, adding auxiliary variables with lower correlations will typically have little incremental benefit, especially when these additional variables are correlated with the auxiliary variables already being used.

Good candidates for auxiliary variables are the same variables used in the analytic model, but measured at different waves. For example, in a test of a program's effect on smoking at wave 4, with smoking at wave 1 as a covariate, smoking at waves 2 and 3 is not being used in the analysis model but should be rather highly correlated with smoking at wave 4. These two variables would make excellent auxiliary variables.

A related strategy is to impute variables that are relatively highly correlated with one another. Suppose, for example, that one would like to conduct program effect analyses on a large number of DVs, say 30. But including 30 DVs, 30 wave-1 covariates, and the corresponding 60 variables from waves 2 and 3 (as auxiliary variables) would be too many, especially considering that there are perhaps 10 background variables as covariates and perhaps 20 dummy variables representing cluster membership. All these hypothetical variables total 150 variables in the imputation model.

In this instance, a principal components analysis on the 30 DVs could identify three or four sets of items that are relatively highly correlated. This approach does not need to be precise; it is simply used as a device for identifying variables that are more correlated. These groupings of variables would then be imputed together. By conducting imputation analyses and analyses in these groupings, nearly all of the benefits of the auxiliary variables would be gained at a minimum cost of time for imputation and analysis.

Practicalities of Measurement: Planned Missing Data Designs

In this section, I outline the developing area of planned missingness designs. If you are following my advice, you are already (or soon will be) using the recommended MI and ML missing data procedures. Thus, it is time to explore new possibilities with respect to data collection.

3-Form design. Measurement is the cornerstone of science. Without measurement, there is no science. When researchers design the measurement component of their study, they are universally faced with the dilemma of wanting to ask more questions than their study participants are willing to answer, given what they are being paid. Invariably, researchers are faced with the choice of asking fewer questions or paying their participants more. The 3-form design (Graham et al. 2006), which has been in use since the early 1980s (e.g., see Graham et al. 1984), gives researchers a third choice. In its generic form, the 3-form design allows researchers to increase by 33% the number of questions for which data are collected without changing the number of questions asked of each respondent. The trick is to divide all questions asked into four items sets. The X set, which contains questions most central to the study outcomes, is asked of everyone. But the A, B, and C sets of items are rotated, such that one set is omitted from each of the three forms. **Table 3** describes the basic idea of the design (Raghunathan & Grizzle 1995 suggested a similar design involving all possible two-set combinations of five items sets).

The benefit of the 3-form design is that one gathers data for 33% more questions than can be asked of any one participant. Also, importantly, at least one-third of the participants provide data for every pair of questions. That is, all correlations are estimable. This feature is not shared by most other measurement designs of this general sort (generically described as matrix sampling). The drawback to the 3-form design is that some correlations, because they are based on only one-third of the sample, are tested with lower power. However, as Graham et al. (2006) show, virtually all of the possible drawbacks are under the researcher's control and can generally be avoided.

Table 3 3-form design

	Respondent received item set?			
Form	X	A	B	C
1	Yes	Yes	Yes	No
2	Yes	Yes	No	Yes
3	Yes	No	Yes	Yes

Two-method measurement. Graham et al. (2006) also described a planned missingness design called two-method measurement (also see Allison & Hauser 1991, who describe a related design). The two-method measurement design stems from the need to obtain good, valid measures of the main DV. Researchers in many domains face a dilemma: (*a*) collect data from everyone with a relatively inexpensive, but questionably valid measure (e.g., a self-administered, self-report measure of recent physical activity), or (*b*) collect data from a small proportion of study participants using a more valid, but much more expensive, measure (e.g., using an expensive accelerometer). The two-method measurement design allows the collection of both kinds of data: complete data for the less expensive measure, and partial data (on a random sample of participants) for the expensive measure. SEM models are then tested in which the two kinds of data are both used as indicators of a latent variable of the construct of interest (e.g., recent physical activity). If certain assumptions are met (and they are commonly met), then this SEM approach allows for the test of the main study hypotheses with more statistical power than is possible with the expensive measure alone and with more construct validity than is possible with the inexpensive measure alone. For details on this design, see Graham et al. (2006).

ATTRITION AND MNAR MISSINGNESS

The methods described up to this point are clear. I believe that the directions we have headed, and the MI and ML methods that have been developed, are the right way to go. My advice continues to be to learn and use these methods. I am certain that the major software writers will continue to help in this regard, and in a very few years, the software solutions will be abundant (e.g., see these recent missing data developments: the feature in Mplus for easing the process of including auxiliary variables; the association between SPSS and Amos software; and the inclusion of MI into SAS/STAT® software). But as we move forward into the arena of attrition and MNAR missingness, the waters get a bit murky. The following section describes more of a work in progress. Nevertheless, I believe that we are making strides, and the way to move forward is becoming clearer.

It has been said that attrition in longitudinal studies is virtually ubiquitous. Although that may be true in large part, also ubiquitous is the fear researchers and critics express that attrition is a threat to the internal validity of the study and the readiness with which researchers are willing to discount study results when more than a few percent of the participants have dropped out along the way. To make matters worse, even the missing data experts often make statements that, if taken out of context, seem to verify that the fears about attrition were well founded. To be fair, the concerns about attrition began well before the missing data revolution, and in the absence of the current knowledge base about missing data, perhaps taking the conservative approach was the right thing to do.

But now we have this knowledge base, and it is time to take another, careful look at attrition. It is important to acknowledge that the conclusions in some studies will be adversely affected by attrition and MNAR missingness, and I am not suggesting that we pretend that isn't the case. But I do believe that the effects of attrition on study conclusions in a general sense are not nearly as severe as commonly feared. In this section, I describe the beginnings of a framework for measuring the extent to which attrition has biasing effects, and I present evidence that the biasing effects of attrition can be tolerably low, even with what is normally considered substantial attrition. Furthermore, I cite research showing that if the recommended safeguards are put in place, the effects of attrition can be further diminished.

With MAR missingness, missing scores at one wave can be predicted from the scores at previous waves. This is true even though the previous scores might be markedly different for stayers and leavers. With MNAR missingness, the missing scores cannot be predicted based on the previous scores; the model that applies to complete cases for describing how scores change over time does not apply to those who have missing data. The practical problem with MNAR missingness is that the missing scores could be anywhere (high or low). And because of this uncertainty, it is possible that clear judgments cannot be made about the study conclusions. The strategies I present in this section serve to reduce that uncertainty. Some of these strategies can be employed after the fact (especially for longitudinal studies), but many of the strategies must be planned in advance for maximum effectiveness.

Some Clarifications About Missing Data Mechanisms

The major three missingness mechanisms are MCAR, MAR, and MNAR. These three kinds of missingness should not be thought of as mutually exclusive categories of missingness, despite the fact that they are often misperceived as such. In particular, MCAR, pure MAR, and pure MNAR really never exist because the pure form of any of these requires almost universally untenable assumptions. The best way to think of all missing data is as a continuum between MAR and MNAR. Because all missingness is MNAR (i.e., not purely MAR), then whether it is MNAR or not should never be the issue. Rather than focusing on whether the MI/ML assumptions are violated, we should answer the question of whether the violation is big enough to matter to any practical extent.

Measuring the Biasing Effects of Attrition

In order for researchers to move toward a missing data approach that focuses on the likely impact of MNAR missingness, they need tools for measuring the effects of missingness (attrition) on estimation bias. Collins et al. (2001) made effective use of the standardized bias (presented as a percent of the standard error) for this purpose: Standardized Bias $=$ $100 \times$ (average parameter estimate $-$ population value)/SE, where $SE =$ the standard error of the estimate, or the standard deviation of the sampling distribution for the parameter estimate in question. Collins et al. (2001) argued that standardized bias greater than 40% (i.e., more than 40% of a standard error) represented estimation bias that was of practical significance. Armed with this tool for judging whether MNAR bias was of practical significance, Collins et al. (2001) showed that the regression coefficient (X predicting Y), where the cause of missingness was Z, was biased to a practical degree only when there was 50% missing on Y and $r_{ZY} = 0.90$. With 50% missing on Y and $r_{ZY} = 0.40$, 25% missing on Y and $r_{ZY} = 0.90$, and 25% missing on Y and $r_{ZY} = 0.40$, the bias was judged not to be of practical significance when the cause of missingness was omitted from the model (i.e., for MNAR missingness). (Note that although Collins et al. 2001 found that the regression coefficient was largely unbiased, the mean of Y did show a practical level of bias in all four of their missingness and r_{ZY} scenarios described above. This may be an issue in some studies.)

The findings of the Collins et al. (2001) study are important for a variety of reasons: (*a*) they demonstrated the usefulness of standardized bias as a way of measuring the practical effects of bias due to attrition, and (*b*) they showed that MNAR missingness alone is often not sufficient to affect the internal validity of an experimental study to any practical extent.

Suppression and inflation bias. Two kinds of attrition bias are inflation and suppression bias. Inflation bias makes a truly ineffective program, or experimental manipulation, appear to be effective. Suppression bias, on the other hand, makes a truly effective program look less effective or an ineffective program look as if it had a harmful effect. When evaluation researchers write about attrition bias, they usually are talking about inflation bias, which is a major concern because it calls into question the internal validity of a study. Suppression bias is much less important if it occurs along with a significant program effect in the desired direction and thus does not undermine the internal validity of the study. On the other hand, suppression bias can be a significant factor if it keeps the truly effective nature of a program from being observed. The possibility of suppression bias is an especially important factor during the planning (and proposal) stages of a project. The chances are reduced that a project will be funded if the power to detect true effects is likely to be diminished unduly because of suppression bias.

Important quantities in describing missingness. In any discussion of missing data and attrition, three quantities are prominently featured: (*a*) the amount of missingness (i.e., percent missing or percent attrition), (*b*) r_{ZY}, the correlation between the cause of missingness, Z, and the model variable containing missingness, Y, and (*c*) r_{ZR}, the correlation between the cause of missingness, Z, and missingness itself, *R*. This last quantity, r_{ZR}, is often manipulated as MAR linear by allowing the probability of a missing Y to be dependent on the quartiles of Z. In Collins et al. (2001), for example, the probabilities of missing on Y were 0.20, 0.40, 0.60, and 0.80 for the first, second, third, and fourth quartiles of Z, respectively. The magnitude of this correlation, r_{ZR}, which depends on the range of these probabilities, can be thought of as the strength of the lever for missingness. That is, a wide range means greater impact on missingness (higher r_{ZR}), and a narrow range means less impact (lower r_{ZR}). Collins et al.

(2001) used a rather strong lever for missingness in their simulations. Their lever for 50% missingness produces $r_{ZR} = 0.447$. A weaker lever for missingness would involve setting the four probabilities for Y missing to different values, for example, 0.35, 0.45, 0.55, and 0.65 for the four quartiles of Z. This lever also produces 50% missingness on Y, but corresponds to $r_{ZR} = 0.224$.

Missing Data Diagnostics

Researchers often want to examine the difference between stayers and leavers on the pretest variables of a longitudinal study. Knowledge will be gained from this practice, but not as much as researchers might think. Because the missingness is not MCAR, any differences observed on pretest variables should not be unexpected. And the information gained cannot indicate the degree to which the missingness is MNAR. Perhaps the best value of this kind of analysis is to identify the pretest variables that most strongly predict missingness later in the study; these variables can be included in the missing data model (e.g., see Heckman 1979, Leigh et al. 1993).

The diagnostics of Hedeker & Gibbons. However, making use of longitudinal data and examining the missingness and the patterns of change over time on the main DV can be very enlightening. Hedeker & Gibbons (1997) provided an excellent example of this kind of diagnostic. In the empirical study they used to illustrate their analytic technique (a pattern mixture model), they plotted the main DV over the four main measurement points in four groups: (*a*) drug group, data for week 6; (*b*) placebo group, data for week 6; (*c*) drug group, data missing for week 6; (*d*) placebo group, data missing for week 6 (where week 6 was the final measure in the study).

These plots (Hedeker & Gibbons 1997, p. 72) show clearly that the changes over time in all four groups were nearly linear. Among those with data for the last time point, the participants in the drug group were clearly doing better than those in the placebo group. Among those who did not have data for the last measure (week 6), the people in the drug condition appeared to be doing even better, and those in the placebo condition appeared to be doing even worse.

Although it is highly speculative to extrapolate from a single pretest to the last posttest, extrapolation to the last posttest makes much more sense when one is working from a clearly established longitudinal trend. That is, there is much less uncertainty about what the missing scores might be on the final measure. In the Hedeker & Gibbons (1997) study, for example, it can be safely assumed that those without data for the final wave continued along the same (or similar) trajectory that could be observed through three of the four time points.

Figure 1 displays the same kind of data from the Adolescent Alcohol Prevention Trial (AAPT; Hansen & Graham 1991). The figure illustrates the same four plots, this time for the program group (Norm) and comparison group (No Norm) for those who did have data for the final follow-up measure at eleventh grade and for those who did not. It is evident that just as in the study described by Hedeker & Gibbons (1997), the plots look reasonably smooth, and it would be a reasonable to assume

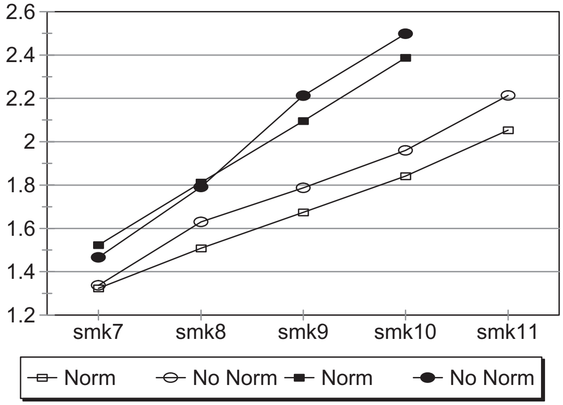

Figure 1

Smoking levels over time for those with and without data for the last wave of measurement. Students in the Norm program group (*square markers*) reported lower levels of cigarette smoking than did students in the Comparison group (*circle markers*) for those who had data for the last wave of measurement (eleventh grade; *white markers*) as well as for those who had missing data for the last wave of measurement (*black markers*).

that the missing smoking scores for eleventh grade followed this same or a similar trajectory.

All of these plotted points make it more difficult to imagine that the missing points are hugely different from where one would guess they would be based on the observed trajectories. On the other hand, it would make little sense to try to predict the missing eleventh-grade data based on the seventh-grade pretest data alone. Based only on those pretest scores, the missing eleventh-grade scores could indeed be anywhere.

The data described by Hedeker & Gibbons (1997) and the data presented here are very well behaved. However, not all plots of the main DV over time will be this smooth. In a longitudinal study, for example, the changes in the DV over time may not conform to a simple, well-defined curve (linear or curvilinear). Under those conditions, predicting the missing final data point would be much more difficult.

Nonignorable (MNAR) Methods

Nonignorable methods (e.g., see Demirtas 2005; Demirtas & Schafer 2003; Hedeker & Gibbons 1997; Little 1993, 1994, 1995) may be very useful. However, it is not necessarily true that any particular method will be better than MAR methods (e.g., normal-model MI or ML) for any particular empirical study. It is well known that methods for handling nonignorable data require the analyst to make assumptions about the model of missingness. If this model is incorrect, the MNAR model may perform even less well than standard MAR methods (e.g., see Demirtas & Schafer 2003).

On the other hand, MNAR methods such as pattern mixture models may, as argued above, be excellent tools for describing the missingness in longitudinal fashion, thereby increasing one's confidence in many instances about the true nature of the missingness. As suggested by Little (1993, 1994, 1995), this type of model can be a good way to perform sensitivity analyses about the model structure. If the same general study conclusions are made over a wide variety of possible missing data models, then one has greater confidence in those study conclusions. In addition, the models suggested by Hedeker & Gibbons (1997) may prove to be especially useful when the longitudinal patterns of the main DV are as smooth as they described.

Strategies for Reducing the Biasing Effects of Attrition

Use auxiliary variables. Probably the single best strategy for reducing bias (and increasing statistical power lost owing to missing data) is to include good auxiliary variables in the missing data model (see Collins et al. 2001). As Little (1995) put it, "one should collect covariates that are useful for predicting missing values." It is important that these variables need only be good for predicting the missing values; they need not be related to missingness, per se. Good candidate variables for auxiliary variables are measures of the main DV that happen not to be in the analysis model. However, if the analysis already involves all measures of the DV (e.g., with latent growth modeling analyses), then the incremental benefit of other potential auxiliary variables is likely to be small.

Collins et al. (2001) showed that including an auxiliary variable with $r_{ZY} = 0.40$ reduced relatively little bias in any of the parameters examined. However, including an auxiliary variable with $r_{ZY} = 0.90$ had a major impact on bias. Later simulations have suggested that the benefit from auxiliary variables begins to be noticeable at about $r_{ZY} = 0.50$ or 0.60. Furthermore, it appears that one or two auxiliary variables with $r_{ZY} = 0.60$ are better than 20 auxiliary variables whose correlations with Y are all less than $r_{ZY} = 0.40$. This is true because such variables are often intercorrelated, and the incremental benefit of adding them to the model is very small.

Longitudinal missing data diagnostics. The excellent strategy described by Hedeker & Gibbons (1997) is discussed above. If longitudinal data are available, this strategy is a good way to describe the missingness patterns. Not all patterns will be good. The best longitudinal patterns are those that reduce the uncertainty about what the missing scores might be at the last waves of measurement.

Measuring intent to drop out. Schafer & Graham (2002) suggested that a potentially good way of reducing attrition bias is to measure participants' intent to drop out of the study. Some people say they will drop out and do drop out, whereas others say they will drop out and do not. Those who do not drop out provide a good basis for imputing the scores of those who do. Demirtas & Schafer (2003) suggested that this approach might be one good way of dealing with MNAR missingness. Leon et al. (2007) performed a simulation that suggested that this approach can be useful.

Collecting follow-up data from those initially missing. Perhaps the best way of dealing with MNAR missingness is to follow up and measure a random sample of those initially missing from the measurement session (e.g., see Glynn et al. 1993, Graham & Donaldson 1993). This strategy is not easy, and it is impossible in some research settings (e.g., where study participants have died). However, even if some studies are conducted that include these follow-up measures from a random sample of those initially missing, it could shed enormous light on the issues surrounding MNAR missingness. With a few well-placed studies of this sort, we would be an excellent position to establish the true bias from using MAR methods and a variety of MNAR methods.

One wrinkle with this approach is that although collecting data from a random sample of those initially missing can be very difficult, collecting data from a nonrandom sample of those initially missing is much easier. Although inferences from this nonrandom sample are generally weaker than inferences possible from a random sample, data such as these may be of much value (e.g., see Glynn et al. 1993).

Suggestions for Reporting About Attrition in an Empirical Study

1. Avoid generic, possibly misleading, statements about the degree to which attrition plagues longitudinal research.
2. Be precise about the amount of attrition; avoid vague terms that connote a missing data problem. Use precise percentage of dropout from treatment and control groups if that is relevant.
3. Missingness on the main DV can be caused by (*a*) the program itself, (*b*) the DV itself, (*c*) the program × DV interaction, or (*d*) any combination of these factors. Perform analyses that lay out as clearly as possible which version of attrition is most likely in the particular study (e.g., see Hedeker & Gibbons 1997).
4. Based on longitudinal diagnostics, assess the degree of estimation bias, for example, using standardized bias (e.g., see Collins et al. 2001) for this configuration and percent of attrition, and determine the kind of bias (suppression or inflation).
5. Draw study conclusions in light of these facts.

Suggestions for Conduct and Reporting of Simulation Studies on Attrition

1. Avoid generic, possibly misleading, statements about the degree to which attrition plagues longitudinal research. Limit the number of assertions about the possible problems associated with attrition. Those of us who do this kind of simulation study must shoulder the responsibility of being precise in how we talk about these topics—what we say can be very influential. Be careful to give proper citations for any statement about the degree to which attrition is known to be a problem or not. Try to focus on the constructiveness of taking proper steps to minimize any biasing effects of attrition.
2. Be precise about the amount of attrition in the simulation study. Provide a

sufficient number of citations to demonstrate that this amount of attrition is of plausible relevance in the substantive area to which the simulation study applies.

3. Be precise about the configuration of attrition simulated in the study. If the configuration of attrition is simulated in a manner different from that used in other simulation studies, then provide a description of the procedure in plain terms and in comparison with the approaches of other simulation researchers (e.g., see the difference in style between Collins et al. 2001 and Leon et al. 2007). Different approaches are all valuable; readers with varying degrees of technical skill just need to know how they relate to one another.

Be precise about the strength of attrition used in the simulation. For example, Collins et al. (2001) specified increasing missingness probabilities for the four quartiles of the variable Z (0.20 for Q1, 0.40 for Q2, 0.60 for Q3, and 0.80 for Q4). As it turns out, this was very strong attrition compared to what it could have been (e.g., 0.35 for Q1, 0.45 for Q2, 0.55 for Q3, and 0.65 for Q4). This is important because the strength used produced bias in the Collins et al. (2001) study that would present practical problems, whereas the latter strength would produce a level of bias that Collins and coworkers would have judged to be acceptably low, even with 50% missingness on Y, and $r_{ZY} = 0.90$ in both cases. Also, present a sufficient number of citations from the empirical literature to demonstrate that the strength of effect used in the simulation actually occurs in empirical research to an extent that makes the study useful. As noted above, the strength of attrition used in the Collins et al. (2001) study was greater than is typically seen in empirical research.

More Research is Needed

Collins et al. (2001) did a nice job of describing the standardized bias concept, which they used as one of the primary yardsticks of practical importance of missing data bias. In their study, they used 40% bias (parameter estimate is four-tenths of a standard error different from the population value) as the cutoff for MNAR bias that would be of practical concern. Anything greater than 40% would be considered to be of practical concern. This implies that anything 40% or less would be considered to be of no practical concern.

Collins et al. (2001) went out on a limb with an estimate of a cutoff for practical effect of MNAR bias, and this study is an excellent starting place for this kind of research. I am reminded of the early days of SEM research, when researchers were struggling to find indices of practical model fit and to find cutoffs for such indices above which a model's fit might be judged as "good." Bentler & Bonett (1980) were the first to provide any kind of cutoff (0.90 for their nonnormed fit index). SEM researchers were eager to employ this 0.90 cutoff, but with considerable experience with this fit index, eventually began to realize that perhaps 0.95 was a better cutoff.

I suggest that researchers involved with work where attrition is a factor (both empirical and simulation studies) begin to develop experience with the standardized bias concept used by Collins et al. (2001). But after years of experience, will we still believe that 40% bias is the best cutoff? It is easy to show that a standardized bias of 40% corresponds to a change in the t-value of 0.4. Can we tolerate such a change? Other issues surround the use of standardized bias. For example, larger sample sizes produce more standardized bias. Future research should address the possibility that different cut points are needed for different sample sizes.

Other indices of the practical impact of attrition bias. In the SEM literature, researchers now enjoy a plethora of indices of practical fit. There are even three or four fundamentally different approaches to these indices of practical fit. We need more such approaches to the practical effects of attrition. I encourage the development of such indices.

Collecting data on a random sample of those initially missing. Many authors have recommended collecting data on a random sample of those initially missing. However, most of this has involved simulation work and not actual data collection. Carefully conducted empirical studies along the lines suggested by Glynn et al. (1993) and Graham & Donaldson (1993) to determine the actual extent of MNAR biases would be valuable, not just to the individual empirical study, but also to the study of attrition in general. If this type of study is conducted properly, it will give us much-needed information about the effect of MNAR processes on estimation bias. It is possible that the kinds of studies for which MNAR biases are greatest are precisely the studies for which collection of additional data on the initial dropouts is most difficult. Nevertheless, even a few such studies will be a great benefit.

Empirical studies testing benefit of intent to drop out questions. The simulation study conducted by Leon et al. (2007) suggests that this strategy is promising. However, empirical studies are needed that make use of these kinds of procedures. Best, perhaps, would be a few carefully conducted studies that examined the combination of this approach with the approach of collecting data on a random sample of those initially missing.

SUMMARY AND CONCLUSIONS

1. Several excellent, useful, and accessible programs exist for performing analysis with missing data with multiple imputation under the normal model and maximum-likelihood (or FIML) methods. Use them! My wish is that 10 years from now, everyone will be making use of these procedures as a matter of course. Having these methods serve as our basic platform will raise the quality of everyone's research.
2. Gain experience with MNAR missing data (e.g., from attrition), especially with the measures of the practical effect of MNAR data. Use the indices that exist, and evaluate them. Come up with your own levels of what constitutes acceptable levels of estimation bias. Where possible, publish articles describing a new approach to evaluating the practical impact of MNAR missingness on study conclusions.
3. Try to move away from the fear of missing data and attrition. Situations will occur in which missing data and attrition will affect your research conclusions in an undesirable way. But don't fear that eventuality. Embrace the knowledge that you will be more confident in your research conclusions, either way. Don't see this possible situation as a reason not to understand missing data issues. Focus instead on the idea that your new knowledge means that when your research conclusions are desirable, you needn't have the fear that you got away with something. Rather, you can go ahead with the cautious optimism that your study really did work.

DISCLOSURE STATEMENT

The author is not aware of any biases that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Allison PD. 1987. Estimation of linear models with incomplete data. In *Sociological Methodology 1987*, ed. C Clogg, pp. 71–103. San Francisco, CA: Jossey-Bass

Allison PD. 2002. *Missing Data*. Thousand Oaks, CA: Sage

Allison PD, Hauser RM. 1991. Reducing bias in estimates of linear models by remeasurement of a random subsample. *Soc. Method Res.* 19:466–92

Arbuckle JL, Wothke W. 1999. *Amos 4.0 User's Guide*. Chicago: Smallwaters

Bentler PM, Bonett DG. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychol. Bull.* 88:588–606

Bernaards CA, Belin TR, Schafer JL. 2007. Robustness of a multivariate normal approximation for imputation of incomplete binary data. *Stat. Med.* 26:1368–82

Collins LM, Schafer JL, Kam CM. 2001. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychol. Methods* 6:330–51

Demirtas H. 2005. Multiple imputation under Bayesianly smoothed pattern-mixture models for nonignorable drop-out. *Stat. Med.* 24:2345–63

Demirtas H, Schafer JL. 2003. On the performance of random-coefficient pattern-mixture models for non-ignorable dropout. *Stat. Med.* 21:1–23

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. R. Stat. Soc.* B39:1–38

Duncan SC, Duncan TE. 1994. Modeling incomplete longitudinal substance use data using latent variable growth curve methodology. *Multivar. Behav. Res.* 29:313–38

Duncan SC, Duncan TE, Hops H. 1996. Analysis of longitudinal data within accelerated longitudinal designs. *Psychol. Methods* 1:236–48

Duncan TE, Duncan SC, Hops H. 1994. The effect of family cohesiveness and peer encouragement on the development of adolescent alcohol use: a cohort-sequential approach to the analysis of longitudinal data. *J. Stud. Alcohol* 55:588–99

du Toit M, du Toit S. 2001. *Interactive LISREL: User's Guide*. Lincolnwood, IL: Sci. Software Intl.

Efron B. 1982. *The Jackknife, the Bootstrap, and Other Resampling Plans*. Philadelphia, PA: Soc. Industrial Appl. Math.

Enders CK. 2001a. A primer on maximum likelihood algorithms available for use with missing data. *Struct. Equ. Model.* 8:128–41

Enders CK. 2001b. The impact of nonnormality on full information maximum-likelihood estimation for structural equation models with missing data. *Psychol. Methods* 6:352–70

Enders CK. 2003. Using the expectation maximization algorithm to estimate coefficient alpha for scales with item-level missing data. *Psychol. Methods* 8:322–37

Glynn RJ, Laird NM, Rubin DB. 1993. Multiple imputation in mixture models for nonignorable nonresponse with followups. *J. Am. Stat. Assoc.* 88:984–93

Graham JW. 2003. Adding missing-data relevant variables to FIML-based structural equation models. *Struct. Equ. Model.* 10:80–100

Graham JW, Cumsille PE, Elek-Fisk E. 2003. Methods for handling missing data. In *Research Methods in Psychology*, ed. JA Schinka, WF Velicer, pp. 87–114. Volume 2 of *Handbook of Psychology*, ed. IB Weiner. New York: Wiley

Graham JW, Donaldson SI. 1993. Evaluating interventions with differential attrition: the importance of nonresponse mechanisms and use of followup data. *J. Appl. Psychol.* 78:119–28

Graham JW, Flay BR, Johnson CA, Hansen WB, Grossman LM, Sobel JL. 1984. Reliability of self-report measures of drug use in prevention research: evaluation of the Project SMART questionnaire via the test-retest reliability matrix. *J. Drug Educ.* 14:175–93

Graham JW, Hofer SM. 1992. *EMCOV Users Guide*. Univ. S. Calif. Unpubl. documentation

Graham JW, Hofer SM. 2000. Multiple imputation in multivariate research. In *Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples*, ed. TD Little, KU Schnabel, J Baumert, 1:201–18. Hillsdale, NJ: Erlbaum

Graham JW, Hofer SM, Donaldson SI, MacKinnon DP, Schafer JL. 1997. Analysis with missing data in prevention research. In *The Science of Prevention: Methodological Advances from Alcohol and Substance Abuse Research*, ed. K Bryant, M Windle, S West, 1:325–66. Washington, DC: Am. Psychol. Assoc.

Graham JW, Hofer SM, MacKinnon DP. 1996. Maximizing the usefulness of data obtained with planned missing value patterns: an application of maximum likelihood procedures. *Multivar. Behav. Res.* 31:197–218

Graham JW, Hofer SM, Piccinin AM. 1994. Analysis with missing data in drug prevention research. In *Advances in Data Analysis for Prevention Intervention Research*, *National Institute on Drug Abuse Research Monograph*, ed. LM Collins, L Seitz, 142:13–63. Washington, DC: Natl. Inst. Drug Abuse

Graham JW, Olchowski AE, Gilreath TD. 2007. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* 8:206–13

Graham JW, Roberts MM, Tatterson JW, Johnston SE. 2002. Data quality in evaluation of an alcohol-related harm prevention program. *Evaluation Rev.* 26:147–89

Graham JW, Schafer JL. 1999. On the performance of multiple imputation for multivariate data with small sample size. In *Statistical Strategies for Small Sample Research*, ed. R Hoyle, 1:1–29. Thousand Oaks, CA: Sage

Graham JW, Taylor BJ, Cumsille PE. 2001. Planned missing data designs in analysis of change. In *New Methods for the Analysis of Change*, ed. LM Collins, A Sayer, 1:335–53. Washington, DC: Am. Psychol. Assoc.

Graham JW, Taylor BJ, Olchowski AE, Cumsille PE. 2006. Planned missing data designs in psychological research. *Psychol. Methods* 11:323–43

Hansen WB, Graham JW. 1991. Preventing alcohol, marijuana, and cigarette use among adolescents: peer pressure resistance training versus establishing conservative norms. *Prev. Med.* 20:414–30

Heckman JJ. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61

Hedeker D, Gibbons RD. 1997. Application of random-effects pattern-mixture models for missing data in longitudinal studies. *Psychol. Methods* 2:64–78

Honaker J, King G, Blackwell M. 2007. *Amelia II: A Program for Missing Data*. Unpubl. users guide. Cambridge, MA: Harvard Univ. **http://gking.harvard.edu/amelia/**

Horton NJ, Kleinman KP. 2007. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.* 61:79–90

Jöreskog KG, Sörbom D. 1996. *LISREL 8 User's Reference Guide*. Chicago: Sci. Software

King G, Honaker J, Joseph A, Scheve K. 2001. Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *Am. Polit. Sci. Rev.* 95:49–69

Lanza ST, Collins LM, Schafer JL, Flaherty BP. 2005. Using data augmentation to obtain standard errors and conduct hypothesis tests in latent class and latent transition analysis. *Psychol. Methods* 10:84–100

Leigh JP, Ward MM, Fries JF. 1993. Reducing attrition bias with an instrumental variable in a regression model: results from a panel of rheumatoid arthritis patients. *Stat. Med.* 12:1005–18

Leon AC, Demirtas H, Hedeker D. 2007. Bias reduction with an adjustment for participants' intent to drop out of a randomized controlled clinical trial. *Clin. Trials* 4:540–47

Little RJA. 1993. Pattern-mixture models for multivariate incomplete data. *J. Am. Stat. Assoc.* 88:125–34

Little RJA. 1994. A class of pattern-mixture models for normal incomplete data. *Biometrika* 81:471–83

Little RJA. 1995. Modeling the drop-out mechanism in repeated-measures studies. *J. Am. Stat. Assoc.* 90:1112–21

Little RJA, Rubin DB. 1987. *Statistical Analysis with Missing Data*. New York: Wiley

Little RJA, Rubin DB. 2002. *Statistical Analysis with Missing Data*. New York: Wiley. 2nd ed.

McArdle JJ. 1994. Structural factor analysis experiments with incomplete data. *Multivar. Behav. Res.* 29(4): 409–54

McArdle JJ, Hamagami F. 1991. Modeling incomplete longitudinal and cross-sectional data using latent growth structural models. *Exp. Aging Res.* 18:145–66

McArdle JJ, Hamagami F. 1992. Modeling incomplete longitudinal data using latent growth structural equation models. In *Best Methods for the Analysis of Change*, ed. L Collins, JL Horn, 1:276–304. Washington, DC: Am. Psychol. Assoc.

Murray DM. 1998. *Design and Analysis of Group-Randomized Trials*. New York: Oxford Univ. Press

Muthén B, Kaplan D, Hollis M. 1987. On structural equation modeling with data that are not missing completely at random. *Psychometrika* 52:431–62

Muthén LK, Muthén BO. 2007. *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén. 4th ed.

Neale MC, Boker SM, Xie G, Maes HH. 1999. *Mx: Statistical Modeling*. Richmond: Virginia Commonwealth Univ. Dept. Psychiatry. 5th ed.

Nunnally JC. 1967. *Psychometric Theory*. New York: McGraw-Hill

Olsen MK, Schafer JL. 2001. A two-part random-effects model for semicontinuous longitudinal data. *J. Am. Stat. Assoc.* 96:730–45

Raghunathan TE. 2004. What do we do with missing data? Some options for analysis of incomplete data. *Annu. Rev. Public Health* 25:99–117

Raghunathan TE, Grizzle J. 1995. A split questionnaire survey design. *J. Am. Stat. Assoc.* 90:54–63

Raudenbush SW, Bryk AS. 2002. *Hierarchical Linear Models*. Thousand Oaks, CA: Sage. 2nd ed.

Rubin DB. 1976. Inference and missing data. *Biometrika* 63:581–92

Rubin DB. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley

SAS Institute. 2000–2004. *SAS 9.1.3 Help and Documentation*. Cary, NC: SAS Inst.

Satorra A, Bentler PM. 1994. Corrections to test statistics and standard errors in covariance structure analysis. In *Latent Variables Analysis: Applications for Developmental Research*, ed. A von Eye, CC Clogg, 1:399–419. Thousand Oaks, CA: Sage

Schafer JL. 1997. *Analysis of Incomplete Multivariate Data*. New York: Chapman & Hall

Schafer JL. 1999. Multiple imputation: a primer. *Stat. Methods Med. Res.* 8:3–15

Schafer JL. 2001. Multiple imputation with PAN. In *New Methods for the Analysis of Change*, ed. LM Collins, AG Sayer, 1:357–77. Washington, DC: Am. Psychol. Assoc.

Schafer JL, Graham JW. 2002. Missing data: our view of the state of the art. *Psychol. Methods* 7:147–77

Schafer JL, Olsen MK. 1998. Multiple imputation for multivariate missing data problems: a data analyst's perspective. *Multivar. Behav. Res.* 33:545–71

Schafer JL, Yucel RM. 2002. Computational strategies for multivariate linear mixed-effects models with missing values. *J. Comput. Graph. Stat.* 11:437–57

Tanner MA, Wong WH. 1987. The calculation of posterior distributions by data augmentation (with discussion). *J. Am. Stat. Assoc.* 82:528–50

von Hippel PT. 2004. Biases in SPSS 12.0 Missing Value Analysis. *Am. Stat.* 58:160–64

Willett JB, Sayer AG. 1994. Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychol. Bull.* 116(2):363–81

Wothke W. 2000. Longitudinal and multigroup modeling with missing data. In *Modeling Longitudinal and Multiple-Group Data: Practical Issues, Applied Approaches, and Specific Examples*, ed. TD Little, KU Schnabel, J Baumert, 1:219–40. Hillsdale, NJ: Erlbaum

Annual Review of Psychology

Volume 60, 2009

Contents

Prefatory

Concepts and Categories

Judgment and Decision Making

Comparative Psychology

Development: Learning, Cognition, and Perception

Early and Middle Childhood

Adulthood and Aging

Substance Abuse Disorders

Therapy for Specific Problems

Adult Clinical Neuropsychology

Child Clinical Neuropsychology

Attitude Structure

Intergroup relations, stigma, stereotyping, prejudice, discrimination

Cultural Influences

Community Psychology

Leadership

Training and Development

Marketing and Consumer Behavior

Psychobiological Mechanisms

Health and Social Systems

Research Methodology

Psychometrics: Analysis of Latent Variables and Hypothetical Constructs

Evaluation

Timely Topics

Indexes

Errata

An online log of corrections to *Annual Review of Psychology* articles may be found at
http://psych.annualreviews.org/errata.shtml