Chapter 2
Analysis of Missing Data

In this chapter, I present older methods for handling missing data. I then turn to the major new approaches for handling missing data. In this chapter, I present methods that make the MAR assumption. Included in this introduction are the EM algorithm for covariance matrices, normal-model multiple imputation (MI), and what I will refer to as FIML (full information maximum likelihood) methods. Before getting to these methods, however, I talk about the goals of analysis.

Goals of Analysis

The goal of any analysis is to obtain unbiased estimates of population parameters. For example, suppose the researcher wants to perform a multiple regression analysis to determine if the variable X has a significant, unique effect on the variable Y, after controlling for the covariate C. The first goal of this analysis is to obtain an estimate of the regression coefficient for X predicting Y that is unbiased, that is, near the population value. The second goal of analysis is to obtain some indication of the precision of the estimate; that is, the researcher wants to obtain standard errors or confidence intervals around the estimate. When these two goals have been achieved, the researcher also hopes to test hypotheses with the maximum statistical power possible. It is in this context that I will talk about the methods for handling missing data. In evaluating the various methods, I will talk about the degree of bias in parameter estimates, and whether or not there is a good way with the strategy for estimating standard errors. Where relevant, I will also evaluate the method with respect to statistical power.

Older Approaches to Handling Missing Data

In this section, I will devote some space to each of these topics: (a) complete cases analysis, (b) pairwise deletion, (c) mean substitution, and (d) regression-based single imputation. With these older methods, the goal is not so much to present a

J.W. Graham, *Missing Data: Analysis and Design*, Statistics for Social and Behavioral Sciences, DOI 10.1007/978-1-4614-4018-5_2,

historical overview of what was typically done prior to 1987. Rather, I want to mention the various approaches, say what is good and bad about them, and in particular, focus on what (if anything) is still useful about them. One thing is clear with these methods, however. None of them were really designed to ***handle*** missing data at all. The word "handle" connotes dealing effectively with something. And certainly none of these methods could be said to deal effectively with missing data. Rather, these methods, usually described as ad hoc, were designed to get past the missing data so that at least some analyses could be done.

Complete Cases Analysis (aka Listwise Deletion)

Complete cases analysis begins with the variables that will be included in the analysis of substantive interest. The analyst then discards any case with missing values on any of the variables selected and proceeds with the analysis using standard methods. The first issue that arises with complete cases analysis relates to whether the subsample on which the analysis is done is a random sample of the sample as a whole. If the missingness is MCAR (see Chap. 1), then the complete cases are representative of the whole, and the results of the analyses will be unbiased. In addition, the standard errors from this analysis are meaningful in the sense that they reasonably reflect the variability around the parameter estimate (although if the estimates are biased, the meaningfulness of these standard errors is questionable).

However, because MCAR missingness is rather a rare occurrence in real-world data, it is almost always the case that cases with complete data for the variables included in the analysis are not representative of the whole sample. For example, in substance abuse prevention studies, it is virtually always true that drug users at the pretest are more likely than nonusers to drop out of the study at a later wave. This means that those with complete cases will be different from those who dropped out. And this difference will lead to estimation bias in several parameters. In particular, means at the posttest will be biased, and Pearson correlations between pretest and posttest variables will be biased.

On the other hand, when missingness is MAR, regression coefficients for pretest variables predicting posttest variables will often be tolerably unbiased. In fact, as noted in Chap. 1, when missingness on Y_2 (Y at time 2) is caused by C_1 (C at time 1; no missing data), then the regression coefficient for X_1 (X at time 1; no missing data) predicting Y_2 is unbiased when C_1 is included as a covariate. In this specific context, complete cases analysis yields b-weights that are identical to those obtained with ML methods (e.g., EM algorithm; Graham and Donaldson 1993).

With respect to bias, complete cases analysis tends to perform quite well, compared to MI and ML analyses, with ANCOVA or multiple regression analysis with several predictors from a pretest, and a single DV from a posttest. And because this type of model is so common, complete cases analysis can often be useful.

However, complete cases analysis fares less well when the proportion of cases lost to missingness is large. Thus, complete case analysis tends to fare much less well with more complex analyses, for example, with a mediation analysis with X, M, and Y coming from three different waves of measurement.

Table 2.1 Hypothetical patterns of missing and observed values

Variable						
A	B	C	D	E	Percent	Data Points
1	1	1	1	1	20	100
1	0	1	1	1	20	100
1	1	0	1	1	20	100
1	1	1	0	1	20	100
1	1	1	1	0	20	100

1 = observed; 0 = missing. 500 total data points

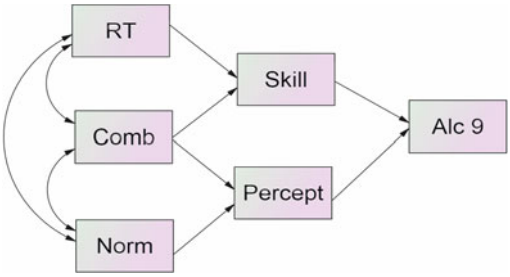

Fig. 2.1 Theoretical mediation model for the adolescent alcohol prevention trial (Hansen and Graham 1991). RT = Resistance Training program vs. control; Norm = Normative Education program vs. control; Comb = Combined (RT + Norm) program vs. control; Skill = behavioral measure of skill in resisting drug use offers; Percept = measure of perceptions of peer drug use; Alc9 = measure of alcohol use at 9th grade

Also, because complete cases analysis involves discarding cases, it often happens that complete cases analysis will test hypotheses with less power. And this loss of power can be substantial if the missingness on different variables in the model comes from nonoverlapping cases. Table 2.1 shows the missingness patterns for such a data set. Although this pattern is somewhat extreme, it illustrates the problem. In this instance, 80 of 500 data points are missing. That is, just 16 % of the total number of data points are missing. However, in this instance, complete cases analysis would discard 80 % of the cases. Discarding 80 % of the cases because 16 % of the values are missing is unacceptable.

In situations such as the one illustrated in Table 2.1, MI or ML methods are clearly a better choice than using complete cases analysis. But even in much less extreme situations, I argue that MI/ML methods are the better choice. In fact, I argue that MI/ML methods are always at least as good as complete cases analysis, and usually MI/ML methods are better, and often they are substantially better than the older methods such as complete cases analysis (Graham 2009).

Graham et al. (1997) compared several different analysis methods with a mediation analysis using data related to the Adolescent Alcohol Prevention Trial (AAPT; Hansen and Graham 1991). A somewhat simplified version of the model tested is shown in Fig. 2.1. The variables on the left in the model represented three program

Table 2.2 Results of analysis of a mediation model based on AAPT data

	Effect	Amos	Mix	EM	CC
RT	→ Skill	.365 (6.29)	.375 (6.36)	.365 (6.98)	.438 (4.56)
Comb	→ Skill	.332 (5.49)	.330 (5.42)	.332 (5.10)	.354 (3.82)
Norm	→ Percept	−.117 (3.31)	−.118 (3.22)	−.117 (3.73)	−.191 (2.31)
Comb	→ Percept	−.270 (7.91)	−.273 (7.89)	−2.70 (8.13)	−.209 (2.90)
Skill	→ Alc9	−.019 (0.48)	−.021 (0.68)	−.019 (0.50)	−.034 (0.62)
Percept	→ Alc9	.143 (4.35)	.119 (3.26)	.143 (3.50)	.135 (1.89)

Note: Table adapted from Graham et al. (1997). Regression coefficients are shown (with corresponding *t*-values shown in parentheses). Amos refers to the Amos Program (Arbuckle 1995); Mix refers to Schafer's (1997) Mix program (multiple imputation for mixed continuous and categorical data); EM refers to the EM algorithm; Standard errors (*t*-values shown) for EM estimates were based on bootstrap methods (Efron 1982); CC refers to complete cases analysis

group variables (variables were dummy coded so that each variable represented a comparison against an information-only control group). The programs were implemented in the seventh grade. The variables in the middle represented the hypothesized mediators of longer term effects. These measures were taken approximately 2 weeks after completion of the programs. The variable on the right represented the longer term outcome (ninth grade alcohol use). In NORM, students received a norms clarification curriculum designed to correct student misperceptions about the prevalence and acceptability of alcohol and other drug use among their peers. In RT, students received resistance skills training. In the COMBined program, students received the essential elements of both NORM and RT curricula. It was hypothesized that receiving the NORM (or COMBined) curriculum would decrease perceptions of peer use, which in turn would decrease ninth grade alcohol use. It was also hypothesized that receiving the RT (or COMBined) curriculum would increase resistance skills, which in turn would decrease ninth grade alcohol use.

Approximately 3,000 seventh grade students received the programs and completed the pretest survey. Approximately the same number completed the immediate posttest survey, which included questions about perceptions of peer use. At the same time as the immediate posttest survey administration, approximately one-third of the students were selected at random to be taken out of the classroom to complete an in-depth, role-play measure of drug resistance skills. Approximately 54 % of those present at the seventh grade pretest also completed the survey at the ninth grade posttest. Given all this, approximately 500 students had data for all measures.

The data described above were analyzed using several procedures, including MI with the MIX program for mixed categorical and continuous data (Schafer 1997), Amos, an SEM program with a FIML feature for handling missing data (Arbuckle 1995), EM algorithm (with bootstrap for standard errors; for example., Graham et al. 1996), and complete cases (CC) analysis. The results of these analyses appear in Table 2.2. The key point to take away from these results is that the results based on complete cases appears to be slightly biased. But more importantly, the mediator → outcome effects were both nonsignificant using complete cases analysis. Thus, had that been our approach, we would not have found significant mediation in this instance (MacKinnon et al. 2002).

Pairwise Deletion

Pairwise deletion is a procedure that focuses on the variance-covariance matrix. Each element of that matrix is estimated from all data available for that element. In concept, pairwise deletion seems like it would be good, because it does make use of all available data. However, because different variances and covariances are based on different subsamples of respondents, parameter estimates may be biased unless missingness is MCAR. In addition, because the different parameters are estimated with different subsamples, it often happens that the matrix is not positive definite, and therefore cannot even be analyzed using most multivariate procedures. An odd by-product of pairwise deletion is that eigenvalues from principal components analysis are either positive (good) or *negative* (bad). With complete cases, eigenvalues are either positive (good) or zero (bad).

In practice, I have found the biggest limitation of pairwise deletion to be the fact that there is no obvious way to estimate standard errors. Estimation of standard errors requires specifying the sample size, and there is no obvious way to do that with pairwise deletion. Thus, with the one exception, outlined in Chap. 8, I do not use pairwise deletion. Even if parameter estimation is all that is needed, better parameter estimates are easily obtained with EM (see below; also see Chaps. 3 and 7).

Mean Substitution

Mean substitution is a strategy in which the mean is calculated for the variable based on all cases that have data for that variable. This mean is then used in place of any missing value on that variable.

This is the worst of all possible strategies. Inserting the mean in place of the missing value reduces variance on the variable and plays havoc with covariances and correlations. Also, there is no straightforward way to estimate standard errors. Because of all the problems with this strategy, I believe that using it amounts to nothing more than pretending that no data are missing. I recommend that people should NEVER use this procedure. If you absolutely must pretend that you have no missing data, a much better strategy, and one that is almost as easy to implement, is to impute a single data set from EM parameters (see Chaps. 3 and 7) and use that.

Averaging the Available Variables

This is the situation in which the mean for a scale is calculated based on partial data when the person does not have complete data for all variables making up the scale. I cover this topic thoroughly in Chap. 9 (coauthored by Lee van Horn and Bonnie

Table 2.3 Missing data patterns

X_1	X_2	X_3	Y
1	1	1	1
1	1	1	0

1 = value observed; 0 = value missing

Taylor). But I wanted also to mention here to distinguish it from mean substitution. The idea of using available variables to calculate a scale score is not at all the same as mean substitution. Rather, I think of it as being a variant of regression-based single imputation (see next section). And as such, this strategy, although not perfect, has much better statistical properties.

Regression-Based Single Imputation

With this strategy, one begins by dividing the sample into those with a variable (Y), and those for whom Y is missing, as shown in Table 2.3. One then estimates a regression model in the first group (X_1, X_2, and X_3 predicting Y) and applies that regression equation in the second group.

For example, in the first group, the regression equation is:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + b_3X_3$$

And because all three X variables have data for the second group, the $\hat{Y}$ values are calculable in the second group. These values are the imputed values and are inserted wherever Y is missing.

Conceptually, this is a good way to impute values. It is good in the sense that a great deal of information from the individual is used to predict the missing values. And as I shall show throughout this book (especially see Chap. 11 on auxiliary variables), the higher correlation between the predictors and Y, the better the imputation will be. In fact, this is such a good way to impute values that it forms the heart of the EM algorithm for covariance matrices and normal-model MI procedures.

However, regression-based single imputation is not a great imputation procedure in and of itself. Most importantly, although covariances are estimated without bias with this procedure (when certain conditions are met), variances are too low. It is easy to see why this is. When Y is present, there is always some difference between observed values and the regression line. However, with this imputation approach, the imputed values always fall right on the regression line. It is for this reason that I do not recommend using this approach. The option available within the MVA package in SPSS (even as recent as version 20) for imputing data from the EM solution is this kind of single imputation (von Hippel 2004). I therefore cannot recommend using this imputed data set (however, please see Chaps. 3 and 7 for other options).

Basics of the Recommended Methods

I have often said that the recommended methods for handling missing data fall into two general categories, model-based procedures and data-based procedures. Model-based approaches rewrite the statistical algorithms so as to handle the missing data and estimate parameters all in a single step. Data-based approaches, on the other hand, handle the missing data in one step, and then perform the parameter estimation in a second, distinct, step. The most common of the model-based procedures are the current crop of structural equation modeling (SEM) programs, which use a FIML feature for handling missing data. The most common of the data-based procedures is normal-model MI. However, with the EM algorithm, this distinction gets a little fuzzy (see below). When an EM algorithm is tailored to produce parameter estimates specific to the situation, EM is a model-based approach. However, when the EM algorithm produces more generic output, such as a variance-covariance matrix and vector of means, which is then analyzed in a separate step, it is more like a data-based procedure. The basics of these recommended approaches are presented below.

Full Information Maximum Likelihood (FIML)

The most common of the model-based procedures are the SEM programs that use a FIML feature for handling missing data. As with all model-based approaches, these programs handle the missing data and parameter estimation in a single step. The FIML approach, which has sometimes been referred to as raw-data maximum likelihood, reads in the raw data one case at a time, and maximizes the ML function one case at a time, using whatever information is available for each case (e.g., see Graham and Coffman in press). In the end, combining across the individuals produces an overall estimate of the ML function. All of these SEM/FIML programs provide excellent (ML) parameter estimates for the model being studied and also provide reasonable standard errors, all in one step.

Amos and Other SEM/FIML Programs

Several SEM programs have the FIML feature, including, in alphabetical order, Amos (Arbuckle 2010), EQS 6.1 (Bentler and Wu 1995), LISREL 8.5+ (Jöreskog and Sörbom 2006; also see Mels 2006), Mplus (Muthén and Muthén 2010), Mx (Neale et al. 2003), and SAS (v. 9.2) Proc CALIS. All of these programs allow ML estimation with missing data and provide good standard errors. Amos has the added advantage of being part of the SPSS package. Amos also has the advantage of being exceptionally intuitive and easy to use. For these reasons, and because SPSS users need more missing data tools, I emphasize Amos a little more here. A more detailed discussion of the workings of Amos can be found in Graham et al. (2003; also see Graham et al. in press).

Basics of the EM Algorithm for Covariance Matrices

First, E and M stand for Expectation and Maximization. Also, please understand that it is not quite proper to refer to "the" EM algorithm. There are several EM algorithms. Collins and Wugalter (1992) described one for estimating LTA models, a type of latent class model. Rubin and Thayer (1982) described an EM algorithm for factor analysis. And early versions of the HLM program (Raudenbush and Bryk 2002) also made use of an EM algorithm (Raudenbush et al. 1991). In each case, the EM algorithm is tailored to produce the ML parameter estimates of interest. The version of the EM algorithm I am talking about in this chapter (and throughout this book) is what I refer to as the EM algorithm for covariance matrices.

As with all of these versions, the EM algorithm for covariance matrices first reads in, or calculates the sufficient statistics, the building blocks of the particular analysis being done, and reads out the relevant parameters. In this case, the relevant parameters are a variance-covariance matrix and vector of means. From here on, when I refer to "the EM algorithm," I am speaking of the version that produces a variance-covariance matrix and vector of means.

The EM algorithm is an iterative procedure that goes back and forth between the E-Step and the M-step.

The E-Step

The sufficient statistics for the EM algorithm are sums, sums of squares, and sums of cross products. The program reads in the raw data, and as each case is read in, it updates the sums, sums of squares, and sums of cross products. Where the data point is observed, it is used directly to update these sums. If the data point is missing, however, the best estimate is used in its place. The best estimate of the missing value is the $\hat{Y}$ from a regression equation using all other variables as predictors. For sums, the value is added directly whether it was observed or missing. For sums of squares and sums of cross products, if one or both values were observed, the value is added directly. However, if both values were missing, then the best estimate is added along with a correction term. This correction term is the residual from the regression with all other variables as predictors. Thus, it is like the error variance added to imputed values in multiple imputation (see below).

The M-Step

Once the sums, sums of squares, and sums of cross products have been estimated, the variance-covariance matrix (and vector of means) can simply be calculated. This concludes the first iteration.

From the variance-covariance matrix and means from the first iteration, one can calculate all of the regression equations needed to predict each variable in the model. During the next iteration, these equations are used to update the "best estimate" when the value is missing. After the sums, sums of squares, and sums of cross products have been calculated at this iteration, a new variance-covariance matrix and vector of means are calculated, and new regression equations are estimated for the next iteration.

This process continues until the variances, covariances, and means change so little from iteration to iteration that they are considered to have stopped changing. That is, when this happens, EM is said to have ***converged***.

The variance-covariance matrix and vector of means from the last iteration are ML estimates of these quantities. Any analysis that requires only a variance-covariance matrix and vector means as input can be used with these EM estimates as input. If the new analysis is something that is simply calculated based on the input matrix, for example, a multiple regression analysis, then those estimates are also ML (note, e.g., that the EM and Amos parameter estimates from Table 2.2 are identical). However, if the analysis itself is an iterative procedure, such as a latent-variable regression model, then the estimates based on the EM variance-covariance matrix and means will be unbiased and efficient but technically will not be ML.

Standard Errors

The one drawback with EM is that standard errors are not produced as a by-product of the parameter estimation. There are other approaches (e.g., see Yuan and Bentler 2000), but the most common approach to estimating standard errors with EM estimates is to use bootstrap procedures (e.g., Graham et al. 1997). Note that the *t*-values based on bootstrapping in Table 2.2 are reasonable, but are somewhat different from the those based on FIML and MI analysis. Although the EM + bootstrapping process is generally more time consuming than FIML or MI, one notable advantage of EM with bootstrapping is that this is a good approach when data are not normally distributed. In this instance, bootstrapping to yield direct estimates of the confidence intervals (which requires one or two thousand bootstraps) provides better coverage than does either MI or FIML with regular (i.e., not robust) standard errors.

Implementations of the EM Algorithm

The EM algorithm for covariance matrices now has many implementations, including SAS Proc MI, Norm (Schafer 1997), and EMCOV (Graham and Hofer 1991). SPSS does have an EM algorithm routine within its MVA module. This is a stand-alone routine that does not interface with other parts of SPSS, but it can be very useful for estimating EM means, variances, and correlations. The latest versions of STATA also have EM capabilities.

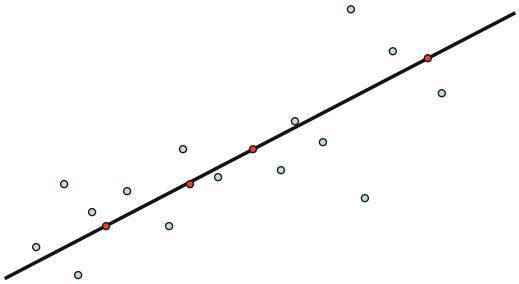

Fig. 2.2 A bivariate distribution with the best-fitting straight line. Imputed values based on regression-based single imputation lie right on the regression line. Real (observed) data points deviate by some amount from the regression line

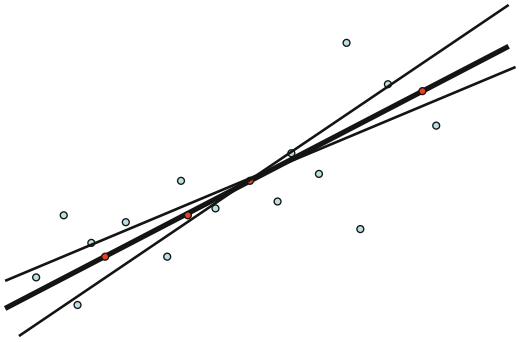

Fig. 2.3 Regression lines are slightly different for different random draws from the population

Basics of Normal-Model Multiple Imputation

In a previous section, I said that the regression-based, single imputation procedure formed the heart of EM and normal-model MI. I also said that regression-based single imputation underestimates variances. That is, there is too little variability in the imputed values. The first reason for this is that the imputed values have too little error variance. This problem is depicted in Fig. 2.2. The observed data points deviate from the regression line by some amount, but, of course, the imputed values lie right on the regression line. This problem is easily resolved simply by adding random normal error to each imputed value (this corresponds to adding the correction term in the E-step of the EM algorithm, as described above).[1]

The second reason there is too little variability relates to the fact that the regression equations used in single imputation are based on a single sample drawn from the population. As depicted in Fig. 2.3, there should be additional variability around

[1] It is this random error that is missing from the data set imputed from the EM solution in the MVA module of SPSS (von Hippel 2004; this remains the case at least through version 20).

the regression line itself to reflect what would occur if there were a different random draw from the population for each imputed data set. Of course, researchers seldom have the luxury of being able to make several random draws from the population of interest. However, bootstrap procedures (Efron 1982) can be used in this context. Or random draws from the population can be simulated using Bayesian procedures, such as Markov-Chain Monte Carlo (MCMC) or data augmentation (Tanner and Wong 1987; Schafer 1997) procedures.

It has been said that data augmentation (DA), which is used in Schafer's (1997) NORM program, is like a stochastic version of EM. DA is also a two-step, iterative process. In the I-step (imputation step), the data are simulated based on the current parameter values. In the P-step (posterior step), the parameters are simulated from the current data. On the other hand, DA converges in a way that is rather different from how EM converges. Whereas EM converges when the parameter estimates stop changing, DA converges when the distribution of parameter estimates stabilizes.

Recall that DA is used in order to simulate random draws from the population. However, as with all Markov Chain procedures, all information at one iteration comes from the previous iteration. Thus the parameter estimates (and imputed data) from two consecutive steps of DA are much more like one another than if they had come from two random draws from the population. However, after some large number of DA steps from some starting point, the parameter estimates are like two random draws from the population. The question is how many DA steps between imputed data sets is enough? The answer (described in more detail in Chaps. 3 and 7) is that the number of iterations for EM convergence is a good estimate of the number of DA steps one should use between imputed data sets.

The Process of Doing MI

Analysis with MI is a three-step process. First, one imputes the data, generating m imputed data sets. With each data set, a different imputed value replaces each missing value. Early writers suggested that very few imputed data sets were required. However, more recent work has suggested that more imputations (e.g., $m = 20$–40 or more) are required to achieve the statistical power of equivalent ML procedures (Graham et al. 2007; see below for more details). The details for performing MI are given in Chaps. 3 and 7.

Second, one analyzes the m data sets with usual, complete data, procedures (e.g., with SAS, SPSS, HLM, etc.), saving the parameter estimates and standard errors from analysis of each data set. Details for performing analyses are given in Chaps. 4, 5, 6, and 7.

Third, one combines the results to get ***MI inference***. Following what are commonly known as Rubin's rules (Rubin 1987), the two most important quantities for MI inference are the point estimate of the parameters of interest and the MI-based standard errors. These and other important quantities from the MI inference process are described below.

Point Estimate of the Parameter

The point estimate for each parameter is simply the arithmetic average of that parameter estimate (e.g., a regression coefficient) over the *m* imputed data sets. It is this average for each parameter of interest that is reported in the article you are writing.

Standard Errors and *t*-Values

The MI inference standard error (SE) is in two parts. One part, *within-imputation variance* (*U*) reflects the regular kind of sampling variability found in all analyses. The other part, between-imputation variance (*B*), reflects the added variability, or uncertainty, that is due to missing data. The within-imputation variance is simply the average of the squared SE over the analyses from the *m* imputed data sets, that is,

$$\mathrm{U} = \sum \mathrm{SE}^2 / m,$$

for each parameter being studied. The between-imputation variance is the sample variance of the parameter estimate (e.g., a regression coefficient) over the *m* imputed data sets,

$$B = \mathrm{S}^2_{\mathrm{P}},$$

where P is the parameter being studied. The total variance is a weighted sum of the two kinds of variance,

$$T = U + (1 + 1/m)B.$$

It should be clear that *B* is the variance that is due to missing data. If there were no missing data, then the variance of the parameter over the *m* imputed data sets would be 0 and the *B* component of variance would be 0. The MI inference standard error is simply the square root of *T*.

Degrees of Freedom (df)

The *df* associated with the *t*-value in Rubin's rules, adapted from Schafer (1997), is

$$df = (m-1) = \left[1 + \frac{\mathrm{U}}{(1+m^{-1})B}\right]^2.$$

The *df* in MI analysis is different from *df* in other statistical contexts; for example, it has nothing to do with *N*. Just looking at the formula for *df* can give insights into its meaning. First, if there were very little missing data, *B* would be very small. At the limit, *B* would tend toward 0, and *df* would tend toward infinity. On the other

hand, if there were much missing data and uncertainty of estimation due to missing data, then *B* would tend to be large in comparison to *U*, and the right-hand term in the brackets would tend to be very small. In that case, *df* would tend toward its lower limit ($m-1$). More conceptually, I think of *df* as indicating the stability of the MI estimates. If *df* is large, compared to *m*, then the MI estimates have stabilized and can be trusted. However, if *df* is small, for example, near the lower limit, it indicates that the MI estimates have not stabilized, and more imputations should be used.

Fraction of Missing Information

The fraction of missing information (*FMI*) in Rubin's rules, adapted from Schafer (1997), is

$$FMI = \frac{r + 2/(df + 3)}{r + 1}$$

where

$$r = \frac{(1 + m^{-1})B}{U}.$$

FMI is an interesting quantity. Conceptually, it represents the amount of *information* that is missing from a parameter estimate because of the missing data. In its simplest form, the *FMI* is theoretically the same as the amount of missing data. For example, with a simple situation of two variables, X and Y, where X is always observed, and Y is missing, say 50 % of the time, $FMI = .50$ for b_{YX}, the regression coefficient for X predicting Y. However, when there are other variables in the model that are correlated with Y, *FMI* will theoretically be reduced, because, to the extent that those other variables are correlated with Y, some of the lost information is restored (see Chap. 11 for a detailed presentation of this issue).

It is important to note that in any analysis, the estimated FMI will differ from the hypothetical value. *FMI* based on the formulas given above is only an estimate. And it can be a rather bad estimate, especially when *m* is small. I do not trust the *FMI* estimate at all unless $m = 40$ or greater. And even then, although I do look at the *FMI* to get a sense of its magnitude, I always bear in mind that the true *FMI* could be a bit different.

What Analyses Work with MI?

It should be clear from reading this book that I believe strongly that normal-model MI is an exceptionally useful analytic tool. Normal-model MI, which is just one of the MI models that has been described in the literature, is (a) without doubt the best

implemented of the available programs, and (b) able to handle an exceptionally wide array of analytic problems.

I think it helps to know that normal-model MI "preserves," that is, estimates without bias, means, variances, covariances, and related quantities. It does not, however, give unbiased estimates for the proportion of people who give a particular answer to a variable with more than two response levels. For example, normal-model MI (and the related EM algorithm) typically cannot be used to estimate the proportion of respondents who respond "none" to a variable asking about the number of cigarettes smoked in the person's lifetime. That proportion is really a categorical quantity, and unless the variable happens to be normally distributed, normal-model MI will get it wrong. The good news is that variables such as this can often be recast so that normal-model MI can handle them. With the lifetime cigarette smoking question, for example, if the variable were recoded to take on the values 0 (never smoked) and 1 (ever smoked), then normal-model MI (and EM) will produce unbiased estimates of the proportion. The reason it works in this instance is that the proportion of people responding "1" is the same as the mean for that recoded version of the variable, and the estimate of the mean is unbiased with normal-model MI.

Normal-Model MI with Categorical Variables

Normal-model MI does not deal with categorical variables with more than two levels, unless they are first dummy coded; any categorical variable with p levels must be recoded into $p-1$ dummy variables. When such variables have no missing data, that is all that needs to be done. When such variables have missing data, the values may be imputed with normal-model MI, but a minor ad hoc fix may be needed for certain patterns of imputed values (Allison 2002; also see Chaps. 3 and 7). Normal-model MI may also be used for cluster data (e.g., students within schools). I discuss this topic in much greater detail in Chap. 6, but suffice it to say here that normal-model MI does just ok with cluster data, and in this instance, other MI models (e.g., Schafer's PAN program; Schafer 2001; Schafer and Yucel 2002) are preferred.

Normal-Model MI with Longitudinal Data

Schafer's (2001) PAN program was developed initially to handle the special longitudinal data problem depicted in Table 2.4. The data came from the AAPT study (Hansen and Graham 1991). The three variables shown were Alcohol (alcohol use), Posatt (beliefs about the positive social consequences of drinking alcohol), and Negatt (beliefs about the negative consequences of drinking alcohol). As shown in the table, students were asked about their alcohol consumption in each grade from fifth to tenth grades. However, the Posatt questions were not asked in eighth grade,

Table 2.4 "Special" longitudinal missing data patterns

	Grade					
	5	6	7	8	9	10
Alcohol	1	1	1	1	1	1
Posatt	1	1	1	0	1	1
Negatt	1	1	1	0	0	0

1 = data observed; 0 = data missing. Data for each case would normally appear in one long row

Table 2.5 Typical longitudinal missing data patterns

	Alcohol in grade						
Pattern	5	6	7	8	9	10	N
1	1	1	1	1	1	1	500
2	1	1	1	1	1	0	200
3	1	1	1	0	1	1	100
4	1	1	1	0	0	0	200

1 = data observed; 0 = data missing

but reappeared on the survey in ninth and tenth grades. The Negatt questions were not asked in any of the last three grades.

Normal-model MI cannot impute data such as those shown, because data for each case (which would normally appear in one long row) would be missing data for Posatt at eighth grade and Negatt at eighth, ninth, and tenth grades. However, PAN (short for panel) adds a longitudinal component (essentially a growth model) to the imputation procedure. Thus, Posatt data for fifth, sixth, seventh, ninth, and tenth grades can be used to make a good guess about the missing value for Posatt at eighth grade. Also, Negatt at fifth, sixth, and seventh grades can be used to make guesses about the missing Negatt scores for eighth, ninth, and tenth grades. Of course, we would be much more confident about imputing the missing Posatt score at eighth grade than we would about imputing the missing Negatt scores. But the point is that this kind of imputation is possible with PAN.

Many people believe, incorrectly, that programs such as PAN must also be used to impute longitudinal data under what I would refer to as typical circumstances (e.g., the pattern depicted in Table 2.5). The data shown in Table 2.5 differ significantly from the data shown in the previous example. With the data shown in Table 2.5, some people have complete data. More importantly, with these data, at least some cases have data for every variable and for every pair of variables. Under these circumstances, longitudinal models, for example, growth models, may be estimated based on a variance-covariance matrix and vector of means (e.g., using SEM procedures; see Willett and Sayer 1994). And because variances, covariances, and means are estimated without bias with normal-model MI (and the corresponding EM algorithm), these normal-model procedures are sufficient for imputing data in this longitudinal context.

Imputation for Statistical Interactions: The Imputation Model Must Be at Least as Complex as the Analysis Model

Researchers are often interested in statistical interactions (e.g., see Aiken and West 1991; Jaccard and Turrisi 2003). One way to form a statistical interaction is simply to obtain a product of two variables, for example,

$$AB = A \times B$$

Suppose one is interested in testing a regression model with main effects and interaction terms, for example, A, B, and AB as predictors of Y. People often ask if they can go ahead and test this kind of interaction model if they did not address the interaction during imputation. The general answer is that the imputation model should be at least as complex as the analysis model. One way of thinking of this is that any variable that is used in the analysis model must also be included in the imputation model.[2]

If a variable is omitted from the imputation model, then imputation is carried out under the model in which the omitted variable is correlated $r=0$ with all of the variables included in the model. Thus, to the extent that there is missing data, the correlation between the omitted variable and any included variable will be suppressed, that is biased, toward zero. Interactions (product of two variables) are commonly omitted from the imputation model. And because the product is a nonlinear combination of two variables, it cannot simply be calculated after imputation. One solution, then, is to anticipate any interactions, and to include the appropriate products in the imputation model.

A more convenient approach to imputation with interactions is available for some classes of variables – categorical variables that fall naturally into a small number (e.g., just two or three) groups. This approach follows from the idea that interactions can also be conceived of as a correlation between two variables (e.g., r_{AY}) being different when some categorical third variable, B, is 0 or 1. With this approach, one simply imputes separately at the two (or more) levels of the categorical variable. The good news is that imputing in this manner allows one, after imputation, to test any interaction involving the categorical variable. For example, if one imputes separately for males and females, then any interaction involving gender can be tested appropriately after imputation. This strategy also works well for treatment membership variables. If one imputes separately within treatment and control groups, then any interaction involving that treatment membership variable can be tested appropriately after imputation.

[2] However, it is acceptable if variables are included in the imputation model that are not included in the analysis model.

Normal-Model MI with ANOVA

The kind of analysis that works best with MI is the kind of analysis that produces a parameter estimate and standard error. Thus, virtually all analyses in the large family of regression analyses lend themselves very well to normal-model MI. Analyses that do not work so well are ANOVA-related analyses, specifically, analyses that focus on sums of squares, F-tests, and the like. Fortunately, it is generally possible to recast problems that are typically handled with some version of ANOVA into some kind of regression analysis.

Analyses for Which MI Is not Necessary

Some analyses do not require the overhead associated with MI. For example, as I outline in Chaps. 4, 5, and 7, analyses (e.g., coefficient alpha analysis or exploratory factor analysis) that do not require hypothesis testing are more readily handled directly by analyzing the EM covariance matrix (see Chap. 7), or by imputing a single data set from EM parameters, and analyzing that (see Chaps. 4, 5, and 7).

Similarly, although one would definitely prefer to use MI for multiple regression analysis, certain quantities in those analyses do not necessarily involve hypothesis testing, and can be handled either by analyzing the EM covariance matrix directly (see Chap. 7), or by analyzing the single data set imputed from EM parameters (see Chaps. 4, 5, and 7). For example, standardized b-weights and R^2 values can theoretically be handled with MI. But it is much easier to estimate these quantities using the EM covariance matrix directly or by analyzing a single data set imputed from EM parameters.

Missing Data Theory and the Comparison Between MI/ML and Other Methods

MI and ML methods for handling missing data were designed specifically to achieve unbiased estimation with missing data when the MAR assumption holds. Thus it is not surprising that when compared against older, ad hoc methods (e.g., listwise deletion, pairwise deletion, mean substitution), MI and ML methods yield unbiased parameter estimates. And regardless of whether the assumptions are met or not, MI and ML yield estimates that are at least as good as the older, ad hoc methods (Graham 2009). This does not mean that the MI/ML methods will always be better than, say, complete cases analysis. But they will always be at least as good, usually better, and often very much better than the older methods (Graham 2009).

An important point here is that missing data ***theory*** predicts that MI and ML will be better than the old methods. There are already numerous simulations to

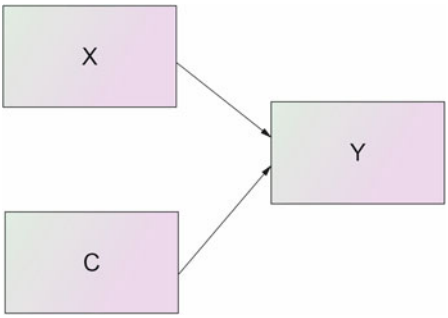

Fig. 2.4 Simple regression model with X and C predicting Y

demonstrate that they are, in fact, better. An important point here is that we do not need more simulations to demonstrate this. What we do need are simulations and other studies that demonstrate the limits of the MI/ML advantage. My recommendation for future research in this area is that the researcher should acknowledge established missing data theory and articulate the reasons why it is either incorrect or incomplete. Here is one example.

Estimation Bias with Multiple Regression Models

Consider the simple regression model depicted in Fig. 2.4. The regression coefficient of X predicting Y (b_{YX}) is of primary interest in the model, and the variable C is included as a covariate. In this instance, X and C are never missing. Y is sometimes missing, and C is the cause of missingness on Y. Graham and Donaldson (1993) demonstrated that under these circumstances, b_{YX} is identical when based on the EM algorithm and on complete cases analysis.

Although this model is a very simple one, it is representative of a very common kind of model. That is, it is common to have a regression model such as that shown in Fig. 2.4, with perhaps several covariates. Even with several covariates, where the pattern of missingness among the covariates could be somewhat complex, complete cases analysis does tend to yield results that are similar to those given by EM and MI. What is important is that regardless of the type of missingness, these EM/MI and complete cases analysis yield similar results for regression coefficients under these circumstances (Graham and Donaldson 1993).

Note that this is not true of other parameters. For example, means and correlations based on complete cases are often substantially biased under the conditions described here. And with more complex models, such as that described in Fig. 2.1 and Table 2.2, the advantage of the MI/ML approach over complete cases analysis can be substantial.

Perhaps the biggest drawback with complete cases analysis is that it is not possible to make use of auxiliary variables. As noted above, with MI/ML methods, the information that is lost to missingness can be partially mitigated by adding auxiliary variables to the model (please see Chap. 11). However, this mitigation makes no sense when there are no missing data, as is the case with complete cases analysis.

Missing Data Theory and the Comparison Between MI and ML

Missing data theory holds that MI and ML are asymptotically equivalent. We do not need new simulations to demonstrate this point. What we do need are studies to define the limits of this equivalence. Under such and such conditions, for example, ML or MI is better.

MI and ML and the Inclusion of Auxiliary Variables

Collins et al. (2001) tested and found substantial support for the following proposition:

> *Proposition 1*. If the user of the ML procedure and the imputer use the same set of input data (same set of variables and observational units), if their models apply equivalent distributional assumptions to the variables and the relationships among them, if the sample size is large, and if the number of imputations, *M*, is sufficiently large, then the results from the ML and MI procedures will be essentially identical (p. 336).

Although their proposition was supported, Collins et al. (2001) noted that it holds in theory. But they also noted that, as typically practiced, MI and ML do have important differences. For example, as practiced, MI users have typically included variables in the imputation model that, although not intended for analysis, were included to "help" with the imputation (we now refer to these variables as auxiliary variables; see Chap. 11).

Users of ML methods, however, in usual practice, are much more likely to limit their models to include only those variables that will be part of the analysis model. Although strategies have been described for including auxiliary variables into some types of ML models (e.g., see Graham 2003, for strategies within a structural equation modeling context), it is not uncommon to see ML models that do not attempt to include auxiliary variables. The exclusion of important auxiliary variables from ML models violates the particulars of the Collins et al. proposition and leads to important differences between the models. An important extension of the concept of including auxiliary variables is that models are less well described for including auxiliary variables in ML approaches to latent class analysis.

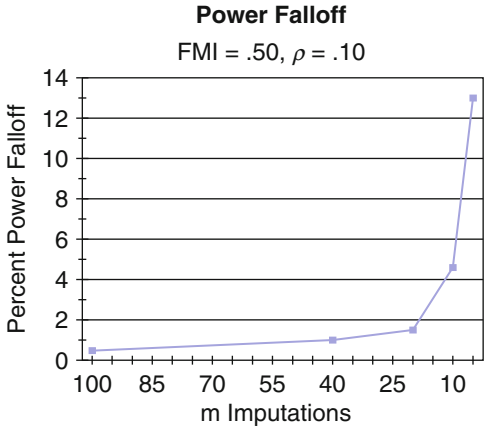

Fig. 2.5 Power falloff with small number of imputations. Power falloff is in comparison to the comparable FIML model. FMI = Fraction of Missing Information; ρ is the population correlation ($\rho = .10$ represents a small effect)

MI Versus ML, the Importance of Number of Imputations

Missing data theorists have often stated that the number of imputations needed in MI in order to achieve efficient estimates was relatively small, and that $m = 3$–5 imputations were often enough. In this context the relative efficiency of the estimate is given by $(1 + \gamma / m)^{-1}$, where γ is the fraction of missing information (Schafer and Olsen 1998). The point made in this context was articulated clearly by Schafer and Olsen:

> Consider … 30 % missing information ($\gamma = .3$), a moderately high rate for many applications. With $m = 5$ imputations, we have already achieved 94 % efficiency. Increasing the number to $m = 10$ raises the efficiency to 97 %, a rather slight gain for doubling of computational effort. In most situations, there is simply little advantage to producing and analyzing more than a few imputed datasets (pp. 548-549).

However, Graham et al. (2007) showed that the effect of number of imputations on statistical power gives a different picture. Graham et al. showed that although the relative efficiency difference might seem small in the context described by Schafer and Olsen (1998), MI with small m could lead to an important falloff in statistical power, compared to the equivalent FIML model, especially with small effect sizes. The numbers below apply to a small effect size in Cohen's (1977) terms ($\rho = .10$). Under these conditions MI with $m = 5$ imputations yields statistical power that is approximately 13 % lower than MI with $m = 100$, and 13 % lower than the comparable FIML analysis (power is .682 for $m = 5$; .791 for $m = 100$; and .793 for the comparable FIML model). Figure 2.5 displays for power falloff compared to FIML for MI with various levels m for $\gamma = .50$. Graham et al. agreed that an acceptable power falloff is a subjective thing, but that power falloff greater than 1 % would be

Table 2.6 Recommended number of imputations needed for power falloff < .01 compared to FIML

γ (Fraction of missing information)				
.1	.3	.5	.7	.9
20	20	40	100	>100

Effect size: $\rho = .10$

considered unacceptable to them. Given this judgment, their recommendations are shown in Table 2.6 for the number of imputations required to maintain a power falloff of less than 1 % compared to FIML.

Computational Effort and Adjustments in Thinking About MI

In the Schafer and Olsen (1998) quote given above, the second to last sentence suggested that the increase in efficiency from .94 to .97 was "… a rather slight gain for doubling of computational effort." It is important to look carefully at this statement. I will go into more detail about this idea in later chapters, but let me say here that their doubling the number of imputations comes nowhere near doubling the computational effort. There are exceptions, of course, but with the latest versions of the common statistical software (especially SAS, but also SPSS to an important extent), and with the automation utilities described in later chapters, it often costs little in additional computational effort to increase from 5 to 40 imputations.

In the earliest days of MI, several factors conspired to make the computational intensity of the procedure undesirable. First, the procedure itself was brand new back in 1987 and was still relatively new even in 1998. Back then, it represented a radically new approach to handling missing data. Second, in 1987 computers were still very slow. With the computers available back then, the difference between 5 and 40 imputations would often have been very important in terms of computational effort. Third, software for performing multiple imputation was not generally available. Certainly, automation features for handling analysis and summary of multiple data sets were not available.

In this context, the MI theorists suggested the impute-once-analyze-many-times strategy. With this strategy, the computational costs of multiple imputation could be amortized over numerous analyses, thereby reducing the overall costs of the MI procedure. Along with the suggestion that perhaps just 3–5 imputations were enough for efficient parameter estimates, MI did not look so bad as an alternative to other possible approaches to handling missing data.

However, as the realities set in for performing multiple imputation in real-world data sets, it became clear that much of the original thinking regarding MI would not be feasible in large-scale research with missing data. In this context, I have begun to realize that the impute-once-analyze-many approach often does not work. Although it would certainly be desirable in many research contexts, it happens far

too often that the researcher needs to make a change in analysis that requires a whole new set of imputations. In addition, pretty much everything is different now. Computers are now very much faster than they were in 1987, and they will continue to get faster. Perhaps more importantly, the software is catching up. The MI feature in SAS (Proc MI; see Chap. 7) is now a highly functional program. And SPSS, with versions 17+, (see Chap. 5) is not far behind. With a fast computer running SAS, it is very feasible to perform multiple imputation separately for each analysis.

References

Aiken, L.S., & West, S.G. (1991). *Multiple regression: Testing and interpreting interactions*. Newbury Park, CA: Sage.

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

Arbuckle, J. L. (1995). *Amos users' guide*. Chicago: Smallwaters.

Arbuckle, J. L. (2010). *IBM SPSS Amos 19 User's Guide*. Crawfordville, FL: Amos Development Corporation.

Bentler, P. M., & Wu, E. J. C. (1995). *EQS for Windows User's Guide*. Encino, CA: Multivariate Software, Inc.

Collins, L. M., Wugalter, S. E. (1992). Latent class models for stage-sequential dynamic latent variables. *Multivariate Behavioral Research, 27,*131–157.

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

Efron, B. (1982). *The jackknife, the bootstrap, and other resampling plans*. Philadelphia: Society for Industrial and Applied Mathematics.

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

Graham, J. W. (2009). Missing data analysis: making it work in the real world. *Annual Review of Psychology*, *60*, 549–576.

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How Many Imputations are Really Needed? Some Practical Clarifications of Multiple Imputation Theory. *Prevention Science, 8*, 206–213.

Graham, J. W., and Coffman, D. L. (in press). Structural Equation Modeling with Missing Data. In R. Hoyle (Ed.), *Handbook of Structural Equation Modeling*. New York: Guilford Press.

Graham, J. W., Cumsille, P. E., and Elek-Fisk, E. (2003). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.). *Research Methods in Psychology* (pp. 87–114). Volume 2 of *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.

Graham, J. W., Cumsille, P. E., and Shevock, A. E. (in press). Methods for handling missing data. In J. A. Schinka & W. F. Velicer (Eds.). *Research Methods in Psychology* (pp. 000–000). Volume 3 of *Handbook of Psychology* (I. B. Weiner, Editor-in-Chief). New York: John Wiley & Sons.

Graham, J. W., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of followup data. *Journal of Applied Psychology*, *78*, 119–128.

Graham, J. W., & Hofer, S. M. (1991). EMCOV.EXE Users Guide. Unpublished manuscript, University of Southern California.

Graham, J. W., Hofer, S.M., Donaldson, S. I., MacKinnon, D.P., & Schafer, J. L. (1997). Analysis with missing data in prevention research. In K. Bryant, M. Windle, & S. West (Eds.), *The science of prevention: methodological advances from alcohol and substance abuse research.* (pp. 325–366). Washington, D.C.: American Psychological Association.

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

Hansen, W. B., & Graham, J. W. (1991). Preventing alcohol, marijuana, and cigarette use among adolescents: Peer pressure resistance training versus establishing conservative norms. *Preventive Medicine*, *20*, 414–430.

Jaccard, J.J. & Turrisi, R. (2003). *Interaction effects in multiple regression*. Newberry Park, CA: Sage Publications.

Jöreskog, K.G. & Sörbom, D. (2006). LISREL 8.8 for Windows [Computer software]. Lincolnwood, IL: Scientific Software International, Inc.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. *Psychological Methods*, *7*(1), 83–104.

Mels, G. (2006) *LISREL for Windows: Getting Started Guide*. Lincolnwood, IL: Scientific Software International, Inc.

Muthén, L. K., & Muthén, B. O. (2010). *Mplus User's Guide. (6th ed.)*. Los Angeles: Author.

Neale, M. C., Boker, S. M., Xie, G., and Maes, H. H. (2003). *Mx: Statistical Modeling*. VCU Box 900126, Richmond, VA 23298: Department of Psychiatry. 6th Edition.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods, Second Edition*. Newbury Park, CA: Sage.

Raudenbush, S. W., Rowan, B., and Kang, S. J. (1991). A multilevel, multivariate model for studying school climate with estimation via the EM algorithm and application to U.S. high-school data. *Journal of Educational Statistics*, *16*, 295–330.

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

Rubin, D. B., & Thayer, D. T. (1982). EM algorithms for ML factor analysis. *Psychometrika*, *47*, 69–76.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. New York: Chapman and Hall.

Schafer, J. L. (2001). Multiple imputation with PAN. In L. M. Collins and A. G. Sayer (Eds.) *New Methods for the Analysis of Change*, ed., (pp. 357–377). Washington, DC: American Psychological Association.

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

Schafer, J. L., and Yucel, R. M. (2002). Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11, 437–457.

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

von Hippel, P. T. (2004). Biases in SPSS 12.0 Missing Value Analysis. *American Statistician*, *58*, 160–164.

Willett, J. B., and Sayer, A. G. (1994). Using covariance structure analysis to detect correlates and predictors of individual change over time. *Psychological Bulletin*, *116(2)*, 363–381.

Yuan, K-H., & Bentler, P.M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology*, *30*, 165–200.