Multiple Imputation For Missing Data: What Is It And How Can I Use It?

Jeffrey C. Wayman, Ph.D.

Center for Social Organization of Schools
Johns Hopkins University

jwayman@csos.jhu.edu
www.csos.jhu.edu

Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago, IL.

Address correspondence and requests for reprints to Jeff Wayman, Center for Social Organization of Schools, Johns Hopkins University, 3003 N. Charles Street, Suite 200, Baltimore, MD 21218. Email jwayman@csos.jhu.edu.

Introduction

Educational researchers have become increasingly aware of the problems and biases which can be caused by missing data. Significant advances have been made in the last 15 years regarding methodologies which handle responses to these problems and biases. Unfortunately, these methodologies are often not available to many researchers for a variety of reasons (e.g., lack of familiarity, computational challenges) and researchers often resort to ad-hoc approaches to handling missing data, ones which may ultimately do more harm than good (Little & Rubin, 1987; Graham, Hofer, Donaldson, MacKinnon, & Schafer, 1997; Schafer & Graham, 2002). There is a need to make available workable methodologies for handling missing data. Multiple imputation is one such method.

Multiple imputation can be used by researchers on many analytic levels. Many research studies have used multiple imputation (e.g., Graham et al., 1997; Wayman, 2002a) and good general reviews on multiple imputation have been published (Graham, Cumsille, & Elek-Fisk, 2003; Graham & Hofer, 2000; Schafer & Olsen, 1998; Sinharay, Stern, & Russell, 2001). However, multiple imputation is not implemented by many researchers who could benefit from it, very possibly because of lack of familiarity with the technique. A paper which provides a more basic computational description than has previously been presented would be a helpful addition to this literature and might invite more researchers to explore and understand the technique. Therefore, the objective of this paper is to help familiarize researchers with the basic process of multiple imputation, including a data example which will guide the reader through the multiple imputation process.

This paper will first present a brief discussion of some missing data issues. Following this will be a description of the workings of the multiple imputation process, with a data example interspersed throughout the description to provide illustration and clarity. Finally, the paper will conclude with a brief discussion of issues surrounding this particular analysis.

Missing Data

Methods for Treatment of Missing Data

The intent of any analysis is to make valid inferences regarding a population of interest. Missing data threatens this goal if it is missing in a manner which makes the sample different than the population from which it was drawn, that is, if the missing data creates a biased sample. Therefore, it is important to respond to a missing data problem in a manner which reflects the population of inference.

It is important to understand that once data are missing, it is impossible not to treat them – once data are missing, any subsequent procedure with that data set represents a response in some form to the missing data problem. As a result, there are many different methods of managing missing data, of which multiple imputation is one. I will present only a brief discussion of missing data methods here before proceeding to the multiple imputation example. More thorough discussion of missing data methods can be found in Graham et al., 2003; Graham & Hofer, 2000; Little and Rubin, 1987; Schafer, 1997; and Schafer and Graham, 2002, to name a few.

Some of the most popular missing data methods involve ad-hoc deletion or replacement of missing data. These methods typically edit missing data to produce a complete data set and are attractive because they are easy to implement. However, researchers have been cautioned against

using these methods because they have been shown to have serious drawbacks (e.g., Little & Schenker, 1995; Graham & Hofer, 2000; Graham et al. 1997; Schafer & Graham, 2002). For example, handling missing data by eliminating cases with missing data ("listwise deletion" or "complete case analysis") will bias results if the remaining cases are not representative of the entire sample. This method is the default in most statistical software. Another common method available in most statistical packages is mean substitution, which replaces missing data with the average of valid data for the variable in question. Because the same value is being substituted for each missing case, this method artificially reduces the variance of the variable in question, in addition to diminishing relationships with other variables. Graham et al. (2003) referred to these traditional methods as" "unacceptable methods." Examples of other unacceptable methods include pairwise deletion and regression-based single imputation.

Additionally, there exist more statistically principled methods of handling missing data which have been shown to perform better than ad-hoc methods (e.g., Little & Rubin, 1987; Graham et al., 1997; Schafer & Graham, 2002). These methods do not concentrate solely on identifying a replacement for a missing value, but on using available information to preserve relationships in the entire data set. Maximum likelihood estimation is one such method. This method requires specification of a statistical model for each analysis and is a sound method for treating missing data, but is often difficult to implement for less-advanced analysts. The Expectation Maximization (EM) algorithm is another method which has been applied to missing data, but obtaining standard errors using EM involves auxiliary methods such as bootstrapping. The topic of this paper, multiple imputation, is a statistically principled method which is more commonly used because of ease of use and available software.

Mechanisms Responsible for Missing Data

Whether implementing multiple imputation or some other method of dealing with missing data, it is important to understand why the data are missing. Graham et al. (2003) described that missing data can informally be thought of as being caused in some combination of three ways: random processes, processes which are measured, and processes which are not measured. Modern missing data methods generally work well for the first two causes, but not for the last. More formally, missing data mechanisms are commonly described as falling into one of three categories, described by Little and Rubin (1987) thusly:

First, data can be "Missing Completely at Random", or MCAR. When data are MCAR, missing cases are no different than non-missing cases, in terms of the analysis being performed. Thus, these cases can be thought of as randomly missing from the data and the only real penalty in failing to account for missing data is loss of power.

Second, data can be missing "Missing at Random", or MAR. In this case, missing data depends on known values and thus is described fully by variables observed in the data set. Accounting for the values which "cause" the missing data will produce unbiased results in an analysis.

Third, data can be missing in an unmeasured fashion, termed "nonignorable" (also called "Missing Not at Random" (MNAR) and "Not Missing at Random" (NMAR)). Since the missing data depends on events or items which the researcher has not measured, this is a damaging situation.

Graham & Donaldson (1993) referred to missing data mechanisms as "accessible" and "inaccessible." An accessible mechanism is one where the cause of missingness can be accounted for. These situations encompass MCAR and most MAR circumstances. An inaccessible mechanism is one where the missing data mechanism cannot be measured. These situations include

nonignorable mechanisms and MAR mechanisms where the cause of missingness is known, but is not measured.

As Graham and Hofer (2000) state, the missing data mechanism is rarely completely inaccessible. Often, the mechanism is actually made up of both accessible and inaccessible factors. Thus, although a researcher may not be confident that the data present a purely accessible mechanism, covering as much of the mechanism possible will usually produce sound results (Graham et al., 1997; Little, 1995; Rubin, 1996). A sensitivity analysis conducted by Graham et al. (1997) showed that the effects of an inaccessible mechanism are often surprisingly minimal in the implementation of multiple imputation. Thus, encountering a situation where a portion of the missing data is inaccessible should not discourage the researcher from applying a statistically principled method. Rather, the attitude should be to account for as much of the mechanism as possible, knowing that these results will likely be better than those produced by naïve methods such as listwise deletion.

A Brief Overview of Multiple Imputation

In multiple imputation, missing values for any variable are predicted using existing values from other variables. The predicted values, called "imputes", are substituted for the missing values, resulting in a full data set called an "imputed data set." This process is performed multiple times, producing multiple imputed data sets (hence the term "multiple imputation"). Standard statistical analysis is carried out on each imputed data set, producing multiple analysis results. These analysis results are then combined to produce one overall analysis.

Multiple imputation accounts for missing data by restoring not only the natural variability in the missing data, but also by incorporating the uncertainty caused by estimating missing data. Maintaining the original variability of the missing data is done by creating imputed values which are based on variables correlated with the missing data and causes of missingness. Uncertainty is accounted for by creating different versions of the missing data and observing the variability between imputed data sets.

It is important to note that imputed values produced from an imputation model are not intended to be "guesses" as to what a particular missing value might be, rather, this modeling is intended to create an imputed data set which maintains the overall variability in the population while preserving relationships with other variables. Thus, in performing multiple imputation, a researcher is interested in preserving important characteristics of the data set as a whole (e.g., means, variances, regression parameters). Creating imputes is merely a mechanism to deliver an analysis which makes use of all possible information.

Why Multiple Imputation?

Multiple imputation is an attractive choice as a solution to missing data problems because it represents a good balance between quality of results and ease of use.

The performance of multiple imputation in a variety of missing data situations has been well-studied and it has been shown to perform favorably (Graham et al., 1997; Graham & Schafer, 1999; Schafer & Graham, 2002). Multiple imputation has been shown to produce unbiased parameter estimates which reflect the uncertainty associated with estimating missing data. Further, multiple imputation has been shown to be robust to departures from normality assumptions and provides adequate results in the presence of low sample size or high rates of missing data.

Multiple imputation also represents a tractable solution to missing data problems. This procedure is computationally simpler than other statistically principled methods such as maximum

likelihood estimation, and as will be shown later, is a method which is intuitive and easy to understand. Although the statistical principles behind multiple imputation are not trivial, user-friendly software exists[1] which employs these procedures such that the researcher can concentrate on learning and implementing the process of multiple imputation rather than the underlying statistics.

Finally, one of the great appeals of multiple imputation is that the required user interaction is familiar to many researchers – multiple imputation produces full, complete data sets on which to perform analyses, and these analyses can be performed by nearly any method or software package the analyst chooses.

The Multiple Imputation Process

Multiple imputation is a surprisingly intuitive procedure. Imagine that you began an analysis with only common sense and a good introductory regression class in your arsenal. Preparing your data for analysis, you notice quite a few missing values. You know that this is commonly dealt with by ignoring cases with missing values, but you're uncomfortable with this because you believe an analysis using only complete cases will produce misleading conclusions. You decide to explore better methods of dealing with the missing data.

Your first idea is to consider ways to substitute values, thus constructing a complete data set. Initially, you think you may try substituting the mean of the variable for each missing data point. This is a good start, but it substitutes the same number for every missing value. In using mean substitution, you see that you not only have artificially reduced the variance of that variable by creating the same value for every missing data point, but your "guess" is not very educated – you haven't utilized any information that other variables could lend.

You move on, thinking you might use a regression line to solve these problems, using predictors that you know are related to the missing data. The regression line produces values which vary from one another and it also lends values which are "educated guesses", or values which use the interrelationships present in the variables. That seems better than mean substitution, but you are uncomfortable with this method also because you feel you are making up data. These values you're plugging in are estimates, not real data, so you think there is variability you're not accounting for, variability due to estimating missing data. But how would you do this?

Suddenly it hits you: since estimates vary, why use just one estimate? Why not produce many estimates of the same data point? What if you could create different versions of your regression line? This would produce different plausible versions of the substituted values, and thus, different plausible versions of how the data might appear in the population. Averaging over these versions would make you more comfortable with your conclusions; observing how these versions vary from one to the other would offer an estimate of the extra variance you are introducing because of missing data estimation.

You have just invented multiple imputation. But how exactly is this implemented in practice? The next few sections will elaborate on the basic premise described above, detailing the three steps needed to implement multiple imputation: First, we must create imputed data sets which are plausible representations of the data. Second, we must perform the chosen statistical analysis on each of these imputed data sets. Third, we must combine the results of these analyses ("average"

[1] For instance SAS, S-Plus, and NORM (Schafer, 1999) are some. Reviews and discussion of software for multiple imputation are given in Hox (1999) and Schafer & Graham (2002).

them) to produce one set of results. The example chosen to illustrate these steps is a simple one, estimating the overall mean of a nationally-administered reading test.

One cautionary note: Schafer and Olsen (1998) point out that like any statistical technique, multiple imputation depends on some assumptions, and responsible use of multiple imputation involves a basic understanding of these assumptions and their implications. Since the goal of this paper is a basic, clear presentation of the process, these assumptions are not discussed here, and many other statistical issues regarding multiple imputation have been intentionally avoided (see Rubin, 1987; Little & Rubin, 2002; Schafer, 1997).

Creating Imputed Data Sets

The first step in multiple imputation is to create values ("imputes") to be substituted for the missing data. In order to create imputed values, we need to identify some model (we'll call it a regression line) which will allow us to create imputes based on other variables in the data set (predictor variables). Since we need to do this multiple times to produce multiply-imputed data sets, we will identify a set of regression lines which are similar to but different from each other.

We can think of these regression lines as presenting different versions of what the actual equation for the missing data might be – plausible, believable regression lines. Producing a few believable versions of the data will allow us to average over these versions later, producing better estimates. Discussion of exactly how this set of regression lines is identified is beyond the scope of this paper; for this paper, we will assume this set is easily produced[2]. The number of imputed data sets to create is up to the analyst. Commonly, researchers choose between 3 and 10 data sets.

Our regression lines will need predictor variables to help preserve relationships in the data. These variables should be chosen either because they are correlated with the missing variable, the reason for missingness, or both. For example, if the missing variable of interest is a high school achievement test score, variables such as the student's previous test scores could be included since they are likely correlated with achievement test scores. Using the same example, suppose students are more likely to be missing the achievement test if they are in the upper grades. Grade in school could then be included in the imputation model as a reason for missingness. Although not detailed here, choosing variables to include in the imputation model is important and the reader is directed to further reference (Collins et al., 2002; Wayman, 2002b).

In this example, the variable of interest is a nationally-administered reading test score, herein referred to as the "national test" and given in normal curve equivalents (NCEs) (see Appendix A for a description of the data). Almost 15% of the data on this test is missing. Four variables were chosen for the imputation model: score on a locally administered reading test ("local test") grade, gender, and special education status (in practice, more variables would be included). There is evidence of missing data bias, as males, special education students, and students who did poorly on the locally-administered test typically did worse on the nationally-administered test. These groups also were more likely to have missing values (See Appendix A). To illustrate the data, Table 1 lists a subset of participant data from this data set.

[2] Methods to create statistical models which impute values are not trivial, and understanding these methods is certainly not unimportant. Since software exists which helps the user with such computations, this discussion will proceed assuming the analyst will have the ability to easily create imputation models. Further reference on this topic can be found in Rubin (1987); Little & Rubin (2002); and Schafer (1997).

Table 1
Selected Data From Full Data Set

Grade	Gender	Special Ed	Local Score	National Score
8	F	no	345	Missing
8	M	no	325	30
8	M	no	308	18
8	M	yes	300	Missing
8	M	no	369	40
8	F	yes	360	10
7	F	no	314	45
7	M	yes	291	Missing
7	F	no	303	10
7	F	no	407	92
7	M	no	375	93
7	F	no	334	Missing
6	F	no	348	56
6	M	yes	383	32
6	F	no	376	60
6	F	no	310	Missing
6	F	no	383	Missing

For this example, we will create three imputed data sets. Imputation was done using the NORM software program (Schafer, 1999), available free at www.stat.psu.edu/~jls/misoftwa.html[3]. The regression lines[4] to impute national test scores from grade, gender, special education, and local test scores are:

(1) National Score = -135.78 + .31(Grade) + 1.14(Male) + -10.68(Special Ed) + .50(Local Score) + error
(2) National Score = -133.40 + .34(Grade) + .81(Male) + -9.96(Special Ed) + .50(Local Score) + error
(3) National Score = -131.51 + .11(Grade) + 1.43(Male) + -10.19(Special Ed) + .49(Local Score) + error

(Note: Male=1, Female=0; Special Ed=1, not Special Ed=0)

If the analyst had to do the imputation by hand, (s)he would begin by identifying cases which were missing the national test. For each of these participants, the analyst would observe that participant's values for grade, gender, special education, and local test, fill these values into the first regression line, add a random error[5], and compute predicted values. These predicted values would be substituted for the missing values to create the first imputed data set. The same procedure would be followed using the second regression line to create the second imputed data set, and also for the third.

[3] Step-by-step tutorials for using NORM are found in Graham et al. (2003) and Schafer and Olsen, (1998).
[4] The regression lines are presented in this paper for illustration purposes. In practice, the analyst rarely sees the imputation models, nor has the need to work with computations from the imputation model. These computations are done by the software and imputations are automatically created.
[5] A random error is added to every imputed value in order to maintain the natural variability in the data set. If error is not added, participants with equivalent predictor values would receive the same imputed value. Commonly, these errors are drawn from a normal distribution.

To illustrate, consider the first student shown in Table 1, an eighth-grade female not in special education who scored 345 on the local test, but is missing the national test. In order to produce an imputed value for this student, we would substitute these values into the first regression equation. We also need to randomly draw an error value to use in this equation (we drew 3.71). Thus, the imputed value for this student is 42.91, figured thusly:

$$42.91 = -135.78 + .31(8) + 1.14(0) + -10.68(0) + .50(345) + 3.71.$$

We ignore the students who are not missing the national test and proceed to the next student missing the national test, an 8th grade male in special education who scored 300 on the local test (see Table 1). Again, we need to randomly draw an error value for the equation; that value is 2.86. The imputed value for this student is 10.02, again figured thusly:

$$10.02 = -135.78 + .31(8) + 1.14(1) + -10.68(1) + .50(300) + 2.86.$$

This procedure would continue for each of the 2894 other students in the data set who are missing the national test score. Once this procedure is finished, each missing score has an imputed value substituted for it, resulting in a fully-complete imputed data set. Note that all of these imputations were created using the first imputation equation, so this procedure has produced imputed data set #1.

Since we have chosen to work with three imputed data sets, we must create two more. To begin creating the second imputed data set, we will follow the same procedure, but this time using Equation 2. Once again, we start with the first student shown in Table 1, the eighth-grade female not in special education who scored 345 on the local test. We randomly draw an error for this imputation (we drew 0.45), resulting in an imputed value of 42.27:

$$42.27 = -133.40 + .34(8) + 0.81(0) + -9.96(0) + .50(345) + 0.45.$$

As before, we would continue to use this equation to impute values for each student missing the national test score, resulting in imputed data set #2. Like imputed data set #1, the second data set is a plausible, but different representation of the population. Imputed data set #3 would then be created using Equation 3, in the same fashion as the first two data sets. Figure 1 illustrates the process of producing imputed data sets.

Figure 1. An Illustration of the Process of Creating Imputed Data Sets.

Grade	Gender	Special Ed	Local Score	National Score
8	F	No	345	????
8	M	No	325	30
8	M	Yes	300	????
7	F	No	314	45
7	M	Yes	291	????
7	F	No	303	10
7	F	No	334	????
6	M	Yes	383	32
6	F	No	376	60
6	F	No	310	????
6	F	No	383	????

National Score = -135.78 + .31(Grade) + 1.14(Gender) + -10.68(Special Ed) + .50(Local Score) + error

Grade	Gender	Special Ed	Local Score	National Score
8	F	No	345	42.91
8	M	No	325	30
8	M	Yes	300	10.02
7	F	No	314	45
7	M	Yes	291	13.26
7	F	No	303	10
7	F	No	334	18.15
6	M	Yes	383	32
6	F	No	376	60
6	F	No	310	27.70
6	F	No	383	53.57

National Score = -131.51 + .11(Grade) + 1.43(Gender) + -10.19(Special Ed) + .49(Local Score) + error

Grade	Gender	Special Ed	Local Score	National Score
8	F	No	345	42.27
8	M	No	325	30
8	M	Yes	300	8.25
7	F	No	314	45
7	M	Yes	291	27.43
7	F	No	303	10
7	F	No	334	38.97
6	M	Yes	383	32
6	F	No	376	60
6	F	No	310	29.18
6	F	No	383	67.13

National Score = -133.40 + .34(Grade) + .81(Gender) + -9.96(Special Ed) + .50(Local Score) + error

Grade	Gender	Special Ed	Local Score	National Score
8	F	No	345	36.23
8	M	No	325	30
8	M	Yes	300	14.38
7	F	No	314	45
7	M	Yes	291	11.09
7	F	No	303	10
7	F	No	334	29.74
6	M	Yes	383	32
6	F	No	376	60
6	F	No	310	23.34
6	F	No	383	55.78

Analyzing Imputed Data Sets

Once the imputed data sets have been created, the analysis of choice is conducted separately for each data set. This analysis can be any analysis you would perform if there were no missing data (e.g., means, regression, ANOVA), in fact, analysis proceeds just like there were no missing data, except the analysis is performed on each imputed data set. Point estimates and variances of these point estimates will be collected from the analyses for combination in the next step.

To illustrate, consider a very simple analysis which computes the overall mean of the national test. For each imputed data set, we will compute a separate mean and variance; these sets of estimates are then saved so they can be combined to produce an overall estimate of the mean and an overall estimate of the standard error (see next section). The results of this analysis are as follows:

Imputed Data Set #1: Mean = 37.8105, Variance = .0187
Imputed Data Set #2: Mean = 37.8488, Variance = .0185
Imputed Data Set #3: Mean = 37.8166, Variance = .0185

Combining Analysis Results

Once the analyses have been completed for each imputed data set, all that remains is to combine these analyses to produce one overall set of estimates. Combining the estimates from the imputed data sets is done using rules established by Rubin (1987). These rules allow the analyst to produce one overall set of estimates like that produced in a non-imputation analysis. In our example, we will use these rules to combine the three sets of means and standard deviations listed above to produce one overall mean and the variance (or standard error) of that mean.

First we will combine the means. Rubin's rules specify that combining the estimates of the parameter of interest (in our example, the mean) is accomplished simply and intuitively by averaging the individual estimates produced by the analysis of each imputed data set. In mathematical terms, this is written generally: $\bar{\theta} = \frac{1}{K}\sum_{k=1}^{K}\hat{\theta}_k$, for K imputed data sets and point estimates $\hat{\theta}_k$ of some parameter of interest θ.

In our example, $K = 3$ and the values of $\hat{\theta}_k$ are the means from each imputed data set. Thus, the estimate of the overall mean is 37.8253, produced simply by averaging the means calculated from the three imputed data sets: $\bar{\theta} = \frac{1}{3}(37.8105 + 37.8488 + 37.8166) = 37.8253$.

Thus, our estimate of the true average national test NCE is about 37.8.

The total variance of $\bar{\theta}$ is equally intuitive, but requires more computation. For K imputed data sets, the total variance of $\bar{\theta}$ is given by the formula $T = \overline{W} + (1 + K^{-1})B$, where $\overline{W} = K^{-1}\sum_{k=1}^{K} W_k$, the average of the K imputed variances, and $B = (K-1)^{-1}\sum_{k=1}^{K}(\hat{\theta}_k - \bar{\theta})^2$. In plain English, this means that the total variance of our estimate is made up of the two components we know we should account for: a component which preserves the natural variability and an additional component which estimates uncertainty caused by missing data. The part of T which measures the natural variability in the data is $\overline{W}$, the "within-imputation" component. This component is analogous to the variance we would produce if we did not need to account for missing data, and is found by merely averaging the variance estimates from each imputed data set. The part of T which measures uncertainty introduced by missing data

involves B, the "between-imputation" component, which measures how the point estimates vary from data set to data set. If the estimates vary greatly from data set to data set, then uncertainty due to imputation is high and *B* is large. If, however, the parameter estimates are all very similar, there is less uncertainty, and *B* is low.

In our example $K = 3$, the $\hat{\theta}_k$'s are the means from the imputed data sets, $\overline{\theta} = 37.8253$ from above, and the values of W_k are the variances from each imputed data set. First, we compute $\overline{W}$ from the formula given above by averaging the variance estimates from the imputed data sets. This produces an estimate of the within-imputation variance: $\overline{W} = 1/3(.0187 + .0185 + .0185) = .0186$. Next, we will compute B in order to estimate the between-imputation component: $B = (3-1)^{-1}((37.8105 - 37.8253)^2 + (37.8488 - 37.8253)^2 + (37.8166 - 37.8253)^2)$. Simplifying, we get $B = (1/2)((-.0148)^2 + (.0235)^2 + (-.0087)^2) = .00042$. Finally, we must substitute our within and between components $\overline{W}$ and B into the formula for the overall variance *T*, so $T = .0186 + (1 + 3^{-1})(.00042) = .0192$. Taking the square root of the variance gives us the standard error of the mean, .1384.

The result of our work is that we have produced an estimate of the overall mean NCE on the national reading test (37.8253) and an estimate of the variance of this mean (.0192). These numbers look the same as if we had not employed multiple imputation and used a more naïve method such as listwise deletion. However, the estimates obtained using multiple imputation represent an attempt to account for sample bias, and we thus believe these estimates are closer to the true population values than, say, estimates using listwise deletion would be.

Discussion

Discussion of Results

In our example, data were missing from a national reading assessment. The assessment and the missing data were thought to be correlated with gender, grade, special education status, score from a local reading test, so these variables were used to help account for missing information from the national test. A basic analysis – computing the overall mean – was undertaken to describe the process of multiple imputation. Given the goals of this paper, the analysis and imputation were understandably rudimentary, but the example still provides good illustration of some important concepts.

As stated earlier, multiple imputation (MI) almost always produces estimates which are more representative of the population than do the more popular methods of handling missing data, listwise deletion (LD) and mean substitution (MS). Means and standard errors were also computed using these methods in order to illustrate earlier points regarding multiple imputation:

Listwise Deletion (LD):	Mean: 38.83, Standard error: 0.146
Mean Substitution (MS):	Mean: 38.83, Standard error: 0.124
Multiple Imputation (MI):	Mean: 37.83, Standard error: 0.138

One criticism of LD and MS is that both produce biased point estimates because both assume that the missing set of participants is similar to the set with valid values. Since males, students in higher grades, students participating in special education, and students with lower local test scores had lower average national test scores, and since these students were more likely to be missing the national test, that assumption is suspect. Given this bias description, we would assume that the sample of students with

valid values would be an artificially higher-scoring sample, thus biasing upward any estimate of mean national test score. In fact, the overall mean calculated using MI is a full NCE lower than that calculated using LD or MS[6]. This is expected, since MI attempts to account for the sample bias described here.

Since MS handles missing data by substituting the same value for each missing data point, standard error estimates from the MS method are necessarily biased downward. This is illustrated in our results. Standard error estimates obtained using MS are 0.022 (15%) and 0.014 (10%) less than the LD and MI estimates, respectively.

It is interesting to note that although the MI method has an extra component estimated in the standard error (the component estimating uncertainty due to missing data) the standard error from MI is still lower than the standard error produced by LD. This is not generally the case, but is not uncommon, and the reasons can be seen by examining the MI variance computation. From the data example, we see that the "within" variance estimate (the average of the variance estimates from the imputed data sets) is 0.186. Taking the square root of this variance gives a standard error of .136, lower than the LD standard error of .146 because the MI method is taking advantage of more participants. The between-imputation component of the overall MI variance is fairly small, so the overall MI standard error results in a smaller number than the LD standard error. In sum, even though best estimate of the population standard error (MI estimate) contains an extra term due to uncertainty, in this example it is still lower than the LD estimate, which suffers from a lower sample size due to attrition.

A final note: the one-NCE difference in parameter estimates between MI and the other methods is not large, but is important because of what it suggests. The simplicity of this example – one of moderate missingness and using only four variables in the imputation model – suggests that performing this analysis in earnest might well uncover a greater degree of missing data bias, not only in the overall estimates, but in relationships among variables. In practice, this analysis would be undertaken with greater care and complexity than that illustrated here. The next section discusses such considerations.

Extending the Example

The aim of this paper was to provide a basic, conceptual overview of how multiple imputation works. Since this example was chosen for ease of illustration, the analysis would be more complicated if it were conducted in actual practice, and there would be further issues for the analyst to consider. Although detailed illustration and discussion of these issues is beyond the scope of this paper, it will be helpful to point out a few of these issues.

It is important to choose a good, inclusive imputation model. The example presented here used only four variables to impute the national reading test, but in practice, there would likely be many more variables included. As discussed earlier, missing data situations usually involve a combination of accessible and inaccessible mechanisms, and the analyst's aim is to account for as much of the accessible mechanism as possible. Thus, the analyst would likely identify an imputation model with more variables, possibly using interactions, and would invest time in deciding which variables were best for this model. In this application, the analyst might concentrate more on variables correlated with the national reading score than variables correlated with missingness – Collins et al. (2002) suggested that in the presence of moderate missing values (less than 25%), variables correlated with the outcome of interest are more impactful than those correlated with missingness.

For illustration purposes, this example used cases which were not missing the imputation variables (grade, gender, special education, local score; see Appendix A), but in reality, these variables also contain

[6] The mean using mean substitution is necessarily equal to the mean using listwise deletion, because mean substitution replaces every missing value with the mean, computed listwise. In this example, mean substitution was accomplished by substituting 38.83 for every missing value.

missing data. In practice, analysts typically impute missing imputation variables along with the missing outcome variables. This practice produces better results than the alternative of ignoring potentially useful predictor variables (Little, 1992, Schafer & Graham, 2002), but the analyst should take care that missingness in the imputation variables is not too high and that these variables themselves are imputed efficiently.

When performing multiple imputation, it is important to assess the quality of the imputation models. The NORM software, for example, uses a Bayesian simulation process called Data Augmentation to draw models (in the example, illustrated as regression lines) to create imputes (Schafer, 1997). This process is assessed using diagnostic tools provided in NORM (Schafer, 1997; Schafer & Olsen, 1998). For example, it is important in Data Augmentation that the models used for imputation be independent of one another; the NORM software provides methods for assessing this.

Summary

In this paper, I have attempted to provide a basic and clear description of the ideas behind and process of multiple imputation. The description offered here is mostly conceptual, aimed at providing a clear understanding of the basic ideas of multiple imputation, a good base from which to learn more about the use of multiple imputation, and an understanding for reading research which uses multiple imputation.

Not long ago, missing data was viewed as something to discard. Researchers knew the bias problems this presented, but there were no methods available to account for missing data bias. Today, however, researchers are not bound by such constraints. Methods such as multiple imputation are available and usable for most researchers, so there is no need to publish studies which suffer from sample bias.

Since missing data have been handled ad-hoc for so long and since multiple imputation and other statistically principled procedures are relatively new, some are skeptical about the validity of their use. Skepticism about any methodology which is unfamiliar and is presented as an improvement upon traditional methodologies is understandable and often necessary. However, it is important to reiterate that the superiority of multiple imputation to traditional methods is based on mathematical fact, not belief or opinion. It is likely that other missing data methods will be developed which are superior to multiple imputation, but until such methods are available, multiple imputation provides a good solution to missing data problems.

There are many issues and aspects regarding multiple imputation which I was unable to discuss here, given the goals and length of the paper, and the user would be well-advised to pursue further learning about multiple imputation before delving into an analysis. Still, I believe this paper provides a good conceptual understanding of multiple imputation and it is my hope this paper can contribute to better research which informs the ultimate goal of everything we do – improving educational outcomes for children.

References

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

Graham, J. W., Cumsille, P. E., & 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., & Donaldson, S. I. (1993). Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. Journal of Applied Psychology, 78, 119-128.

Graham, J. W., & Hofer, S. M. (2000). Multiple imputation in multivariate research. In T. D. Little, K. U. Schnabel, & J. Baumert, (Eds.), Modeling longitudinal and multiple-group data: Practical issues, applied approaches, and specific examples. (201-218). Hillsdale, NJ: Erlbaum.

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. (325-366). Washington, D.C.: American Psychological Association.

Graham, J. W., & Schafer, J. L. (1999). On the performance of multiple imputation for multivariate data with small sample size. In R. Hoyle (Ed.), Statistical Strategies for Small Sample Research. Thousand Oaks, CA: Sage.

Hox, J. J. (1999). A review of current software for handling missing data. Kwantitatieve Methoden, 20 (62), 123-138.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measures studies. Journal of the American Statistical Association, 90 (431), 1112-1121.

Little, R. J. A. (1992) Regression with missing X's: A review. Journal of the American Statistical Association, 87 (420) 1227-1237.

Little, R. J. A., & Rubin, D. B. (2002). Statistical analysis with missing data. New York: John Wiley & Sons.

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

Little, R. J. A., Schenker, N. (1995). Missing data. In G. Arminger, C. C. Clogg, D. B. Sobel (Eds.) Handbook of Statistical Modeling for the Social and Behavioral Sciences. New York, NY: Plenum.

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

Rubin, D. B. (1996). Multiple imputation after 18+ years. Journal of the American Statistical Association, 91 (434), 473-489.

Schafer, J.L. (1999). NORM: Multiple imputation of incomplete multivariate data under a normal model, version 2. Software for Windows 95/98/NT, available from http://www.stat.psu.edu/~jls/misoftwa.html.

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

Schafer, J. L. & Olsen, M. K. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33 (4), 545-571.

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

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

Wayman, J. C. (2002a). The utility of educational resilience for studying degree attainment in school dropouts. Journal of Educational Research, 95 (3), 167-178.

Wayman, J. C. (2002b, April). Practical Considerations in Constructing a Multiple Imputation Model– A Data Example. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Appendix A – Description of Sample

The data for the examples used in this paper come from a large school district in the United States. Variables used for these examples included a participant's grade, gender, participation in special education, normal curve equivalent on a nationally-administered reading test, and raw score on a locally-administered reading test. Local test scores ranged between 232 and 430, but approximately 95% of the data fell between 303 and 383.

The sample was restricted for these examples in order to provide the clearest possible explanation. Participants were included if they were in grades 6, 7, or 8, and were not missing responses for gender, special education status, and local reading test. There were 19373 participants in the resulting sample, 2896 (15%) of whom were missing the national test score. Table 2 describes the sample.

Table 2
Description of Sample

Variable	N	Average NCE for National Test	Percent Missing National Test
Grade			
6	5897 (30%)	39.59	11%
7	7002 (36%)	38.18	17%
8	6474 (33%)	38.79	16%
Special Education			
Yes	3657 (19%)	20.47	22%
No	15716 (81%)	42.68	13%
Gender			
Male	9888 (51%)	36.85	18%
Female	9485 (49%)	40.76	12%
Local Test		$\rho = .67$	
Total	19373	38.83	15%