Imputing Missing Data: A Comparison of Methods for Social Work Researchers

Jeanne A. Saunders, Nancy Morrow-Howell, Edward Spitznagel, Peter Doré, Enola K. Proctor, and Richard Pescarino

Choosing the most appropriate method to handle missing data during analyses is one of the most challenging decisions confronting researchers. Often, missing values are just ignored rather than replaced with a reliable imputation method. Six methods of data imputation were used to replace missing data from two data sets of varying sizes; this article examines the results. Each imputation method is defined, and the pros and cons of its use in social science research are identified. The authors discuss comparisons of descriptive measures and multivariate analyses with the imputed variables and the results of a timed study to determine how long it took to use each imputation method on first and subsequent use. Implications for social work research are suggested.

KEY WORDS: *data analysis; data imputation methods; missing data; research methods*

"Five hundred high school students completed the longitudinal study... The analysis suggests that a significant difference was found between..."

These hypothetical results may appear to be positive, but the researcher failed to report that originally 850 students were in the study, and that each year 5% to 6% of the sample could not be found because they had moved, no longer had a phone, or chose not to participate. Furthermore, because of incomplete data for some variables, researchers had to drop other cases from the analysis. So in reality, more than 50% of the original sample might not be included, or accounted for, in this statement. It is possible that the participants not included in the final analysis have different characteristics from those who were included. How does this dearth of data affect the outcomes reported? Unfortunately, this scenario is all too common in the social work research reported in the literature. This article summarizes the hazards of ignoring missing data and identifies six data imputation methods that can resolve this problem. To examine how results might differ based on the imputation procedure selected, each of these methods was used on two different data sets, each with missing values. The results effectively demonstrate the importance of dealing with missing data and the many issues confronting the social work researcher in this regard.

The researcher's goal is to conduct the most accurate analysis of the data to make valid and efficient inferences about a population to guide practitioners and researchers alike (Schafer & Graham, 2002). Accomplishing this goal requires choosing the most appropriate method to handle missing data. Too often, social work researchers ignore missing data and their effects on data analysis, thus limiting the researcher's ability to achieve this goal. Ignoring missing data typically occurs when there is a widespread failure to understand the significance of the problem or a lack of awareness of the solutions to the problem of missing data (Figueredo, McKnight, McKnight, & Sidani, 2000).

The handling of missing data is not typically addressed in research reports; literature reviews prove this point. Of approximately 100 articles reviewed between 2001 and 2003 from three social work research journals (*Journal of Social Service Research*, *Social Work*, and *Social Work Research*), only 15 percent reported any information about the amount of missing data or how missing data were handled in the analysis. Because virtually all social science survey research involves some incomplete data, treatment of missing data should be a universal concern and addressed in all research reports.

Numerous methods exist to handle the problem of missing data. They include both "old" methods requiring just a few mathematical computations and "new" methods requiring more complex computations that are increasingly easier for social work researchers to perform with statistical programming software. Here we examine the traditional methods, including listwise deletion (the least sophisticated method), mean substitution, hotdecking, and regression imputation. In addition, we discuss two procedures requiring the creation of five imputed, or "implicate," data sets: The first method used one of the implicate data sets (single implicate) and the second, and most complex method, used the average of five implicate data sets (multiple implicate). Each of these methods has its pros and cons, and researchers must also consider the amount of time required to conduct each method of data imputation and the associated analysis. For this study, we conducted each of the imputation methods twice to simulate the initial learning associated with each method (referred to as "first use") and its use on subsequent occasions ("second use").

UNDERSTANDING THE NATURE OF MISSING DATA

Missing data present many challenges for the social work researcher. One challenge is to determine why data are missing. Error on the part of the researcher, those collecting or entering data, and study participants may be to blame. For example, missing data often occur when a participant refuses or forgets to answer a question, when the instrument has skip patterns, or when an interviewer forgets to ask a question.

It is also important for the researcher to identify and report any patterns to the missing data (Schafer & Graham, 2002). Doing so not only helps the consumer of the research to understand the data more completely, but it also justifies the choice of the data imputation method used by the researcher. In addition, data imputation methods, similar to other statistical procedures, are based on assumptions about patterns of missing data. Identifying the patterns of missing data helps the researcher determine whether the missing values are missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR).

Missing completely at random means that the probability of "missingness" of a variable is not related to any of the study variables (Schafer & Graham, 2002; Streiner, 2002). That is, the data are missing due to some totally unrelated event; for example, a client does not finish a depression inventory because a picture in the office fell off the wall and startled her. This type of event occurs so rarely that it is usually best to categorize the missing data as MCAR.

If data are *MAR*, omitted data may be related to at least one other variable in the study but not to the outcome being measured (Schafer & Graham, 2002; Streiner, 2002). For example, an elderly person may have more difficulty getting to an appointment to complete the study questionnaire because of age (a measured variable) but not because of his or her level of depression (the outcome being measured). It is often difficult, however, for researchers to be certain of the relationship between missing data and these variables. Consequently, they may be unable to distinguish data that are truly MAR from data MCAR.

Most frequently, data are MNAR. This means that the reason for the missingness is related to one or more of the outcome variables or that the missingness has a systematic pattern (Pigott, 2001; Schafer & Graham, 2002). For example, participants may drop out of a study on depression and not complete the final questionnaire because they are not seeing improvement from the intervention, or conversely, they may feel so much better that they perceive no more need for the intervention. Also, participants with extreme opinions about an issue are less likely to respond to survey items that ask about that topic (Raaijmakers, 1999). In both of these cases, there is a pattern to the missing values, even if the researcher is not aware of it. Therefore, the best decision about these missing values is that they are MNAR.

There are some relatively easy ways researchers can examine their data to determine whether missing data follow a pattern. During instrument development, response sets should include "don't know," "does not apply," or "refused" responses. These responses allow the researcher to distinguish among these "no answer" responses. For example, if the majority of responses to an item asking for income information are "refused," the researcher would be confident that there was a pattern to the responses. Once data have been collected, researchers can test their data to look for patterns in missing values using regression analysis (Orme & Reis, 1991). This can be accomplished by substituting a constant value (such as the mean) for all missing values on a variable as well as creating a dichotomous variable that indicates whether the value was originally missing or not. By including this newly created variable in the regression equation, the parameter estimate will indicate whether missingness is related to the dependent variable, and thus patterns of missingness can be identified. A third method sometimes used to address the issue of participants dropping out of a study is conducting an attrition analysis to determine whether those who dropped out were significantly different from those who completed the study on certain indicators. Finding no significant difference between the two groups on selected demographic variables is usually taken to mean that there are no differences between the completers and noncompleters, suggesting that the data are MAR. However, even when no differences are found, it is more likely that true differences between the two groups were not **discovered** rather than that differences do not **exist** (Streiner, 2002).

It is important to distinguish among these patterns of missing data (MCAR, MAR, MNAR) for two reasons. The first relates to how representative the observed variables are in relation to the population. Schafer and Graham (2002) noted that when data are MNAR, the sample typically is less like the population it is assumed to represent. Statistically, sample means are more biased and decreased standard deviations are found. The second reason to identify these patterns is that data imputation methods typically assume that, at a minimum, data are MAR. However, most tests have concluded that the more advanced imputation methods, such as multiple implicate, are robust and produce nearly as good results without strictly meeting this assumption (Little & Rubin, 2002).

The percentage of missing data for each of the variables is important to note as well. "Small" percentages of missing values are less problematic and may be corrected with simpler data imputation methods (such as mean substitution) that would present more problems with larger amounts of missing values. There is no consistent definition of "small amount of missing data" in the literature. Rather, it ranges from 5% or less (Tabachnick & Fidell, 1983) to 20% or less of values (Little & Rubin, 2002). It also matters whether the variable for which there is missing data is an independent or dependent variable (Orme & Reis, 1991). In regression analysis, for example, if an independent variable is missing in a large percentage of values, slope estimates are less affected than if the dependent variable has a comparable percentage of missing values. When variables are used in bivariate or multivariate analyses, the amount of data available for the analysis can be seriously attenuated because of missing values in each of the variables (Pigott, 2001). For example, when conducting a correlation with two variables, both missing 5% of their values, the number of cases available for the analysis could be reduced by up to 10% when all cases with missing values are excluded from the analysis. This becomes increasingly important as the sample size of the study decreases and the number of variables used increases.

Another challenge for social work researchers is to understand how missing data affect the statistical analysis of the data. Typically, missing data lead to inaccurate parameter estimates and biased standard errors and population means (Graham, Taylor, & Cumsille, 2001). These numbers are the basis for more advanced statistical calculations. In effect, this means that the researcher may be reporting results that appear to be statistically significant or insignificant when they truly are not.

IMPUTATION METHODS

Many methods are available to help the researcher impute missing values. The choice of imputation method is usually dependent on a continuum of considerations that includes a researcher's knowledge, skills, and available resources. The methods described here fall along that continuum. We discuss each method briefly, weigh its pros and cons, and identify its appropriate use.

Listwise Deletion

The most common—and easiest—method of dealing with missing data is listwise deletion, also called complete-case analysis (Schafer & Graham, 2002). It was included in this study because it is the default method for most computer analysis programs (for example, SAS and SPSS) and consequently is widely used among social work researchers even though it is not always reported.

When listwise deletion is used, the computer program automatically deletes any case that has missing data for any bivariate or multivariate analysis.

Even though each variable may be missing only a small percentage of responses, collectively a large portion of the data may not be used as cases are deleted. This reduction in sample size translates into reduced statistical power and brings into question how representative the remaining sample is of the population being studied. The remaining cases are more likely to be representative if only a few cases were discarded from the analysis. Because of this systematic loss of data with listwise deletion, there is an increased risk of bias if there is any pattern to the missing data—a risk that is lessened only when the data are MCAR (Pigott, 2001; Schafer & Graham, 2002). Some researchers have characterized listwise deletion as the least desirable data imputation method because of these biases and have warned against its use (Graham et al., 2001). Only with a large sample and relatively small amount of missing data may it be appropriate and most expedient to use listwise deletion. This is because there would be little loss of explanatory power, and how representative the sample is would not be brought into question. An advantage of this method is that no extra time is needed for a researcher to conduct the analysis.

Mean Substitution

The second method used in this study, mean substitution, has been described as "archaic" (Graham & Hofer, 2000) but is still used and discussed in the literature. To use this method, the mean of the total sample for a variable is substituted for all of the missing values in that variable. For example, if the average age of the participants in the study is 72.4 years, then 72.4 is used to replace all missing values for age for any case in the data set. Mean substitution is a quick and easy way to recover cases. By using the mean, the estimate of the mean for the variable is not affected. However, this method is based on an assumption that the missing values are MCAR, which as discussed earlier is rarely the case. Furthermore, the estimate of the standard deviation and variance (used in calculating other parametric tests) is reduced, resulting in biased and deflated standard errors (McDonald, Thurston, & Nelson, 2000; Pigott, 2001; Streiner, 2002). The many variations of this method (for example, using the mean of a subsample of the population) share the same advantages and disadvantages of substitution of the sample mean. Although there is some debate about using this method because of the inherent bias that results (Graham & Hofer, 2000; McDonald et al., 2000), it may be appropriate if only a small number of cases are missing values. In terms of researcher time, this method initially takes some time to program but the amount of time is greatly reduced on subsequent uses.

Hotdecking

A third method, *hotdecking*, identifies a person in the data set with complete data who is similar on an identified correlated characteristic to a person with incomplete data and uses that person's score to replace the missing value. To do this, a correlation matrix is used to determine which are the most highly correlated variables (for example, variables Y and Z) with the variable that has missing data (variable X, for example). The data are then sorted by one of these highly correlated variables (Y or Z) from lowest to highest values. Using this sorted data set, the missing values for variable X are replaced by the value that appears for the preceding participant. As a result, missing values are replaced with a value from a case that is similar on a highly correlated variable. For example, participant "Mrs. J.," who did not report her age, is assigned the value from the preceding participant "Mrs. I.," which is 70.1, and participant "Mr. R.," who did not report his age, is assigned the value from the participant preceding his case in the data set "Mr. Q.," which is 73.3. This method works well when the variable used to sort the data is highly predictive of the variable with the missing values and when there is a large sample so that a similar case is easily identified (Streiner, 2002).

Using a similar case is realistic and preserves some of the measurement error that would likely be found if the value had been completed by the respondent (Schafer & Graham, 2002; Streiner, 2002). One of the advantages of hotdecking, compared with mean substitution, is that the standard deviation of the variable with the inserted values better approximates the standard deviation value for the variable without the substituted values. However, standard deviations are still likely to be lower overall (Streiner, 2002). When using hotdecking, bias is more likely to occur in regression equations than when calculating measures of central tendency. Another drawback of hotdecking is its difficulty to implement; programming requires a great deal of time and labor. Even on subsequent uses, time is not reduced as the programming is not easily transferred between data sets. With the greater availability of imputation software, hotdecking is no longer as popular as it once was.

Regression Imputation or Conditional Mean Imputation

The fourth method used in this study, regression imputation (sometimes identified as conditional mean imputation), is a more sophisticated method than the three previously discussed. To begin, several predictors of the variable with missing values are identified using a correlation matrix. The best predictors (that is, those with the highest correlations) are selected and used as independent variables in a regression equation. The variable with missing data is used as the dependent variable. Cases with complete data for the predictor variables are used to generate the regression equation; the equation is then used to predict missing values for incomplete cases. In an iterative process, values for the missing variable are inserted and then all cases are used to predict the dependent variable. These steps are repeated until there is little difference between the predicted values from one step to the next. That is, they converge. The predictors from the last round are the ones that are used to replace the missing values.

Compared with the three other methods discussed, regression imputation uses the most sources of information (across both items and observations) to predict missing values and "theoretically" provides good estimates for missing values (McDonald et al., 2000). However, several disadvantages to using this model have been identified and are usually considered to outweigh the advantages (Graham & Hofer, 2000; Little & Rubin, 2002). First, because the replaced values were predicted from other variables they tend to fit together "too well." That is, they do not reflect random error or variance, and so standard errors are deflated (Allison, 2002). One must also assume that there is a linear relationship between the variables used in the regression equation when there may not be one. These concerns can result in overestimated model statistics and lower significance values, which lead the researcher to falsely report statistical significance. Another disadvantage of this method is that good predictors must be present in the data set. For both regression imputation and hotdecking, replacing the missing values is more challenging and less practical when the highly correlated items also have missing values (Raaijmakers, 1999). One advantage of this method is that software programs are available to conduct this method. Regression imputation was conducted in this study using the Stata (2001) software package. Once learned, the amount of time needed to conduct this method is relatively short.

Implicate Data Sets

To complete the last two imputation methods in this study, using one implicate data set and the average of five implicate data sets, we used the NORM statistical software program (Schafer, 1998) because it was available online at no cost (see http://www.stat.psu.edu/~jls/misoftwa.html). Other software programs, such as SPSS MissingValues (SPSS, Inc., 2002), EMCOV (Graham & Hofer, 1993), and SAS (2000), are available to complete this method of data imputation. All of these programs use similar statistical methods to achieve their results. The NORM program initially required a large amount of time to understand its methods and how to conduct it. However, once learned, the amount of time required on subsequent uses was greatly reduced.

These imputation methods restore the error variability and the variance in the covariance matrix that is lost when using any of the four methods discussed earlier (Graham & Hofer, 2000). This type of imputation is completed by imputing missing values multiple times, creating a complete data set each time. To begin, the researcher identifies the variables with missing values and a smaller number of highly predictive variables. To complete the imputation, NORM uses a complex series of steps based on the expectation maximization (EM) algorithm. Those steps include estimating the missing values with regression imputation, calculating the means and covariance matrix with the imputed values using formulas that account for residual variances and covariances, and then repeating these steps in an iterative process. During this process, a data augmentation procedure is also used that randomly selects starting values from a distribution of observed and imputed values (Streiner, 2002). (For a more detailed description of the EM algorithm, see Allison, 2002; Rubin, 1991). Each iteration creates a new data set. For purposes of this analysis, 1,000 iterations (resulting in 1,000 data sets) were completed.

To use one implicate data set, which is the fifth method used in this study, we chose the 200th imputed data set for use in the analysis. For the sixth method, multiple implicate data sets, five imputed data sets (each 200th iteration) were selected from the 1,000 that were created. Identical analyses (ordinary least squares regression, in our example) were then conducted on each data set and results combined or "rolled up" to produce less biased estimations of the parameters and their standard errors. In our regression example, slope coefficients were averaged across the data sets to produce one set of estimates, and the standard error for each slope was calculated from the five error estimates as well as the variability between the slope estimates (Rubin, 1987).

The final method uses multiple sources of information to predict a missing value so variance is maximally preserved, and by using a different random seed at the start of each imputation pass, variance between the data sets more accurately reflects the uncertainty in imputing missing data. This procedure is considered the mathematical "gold standard" for the most accurate method of data imputation (Little & Rubin, 2002). The rolled up implicate data set method works equally as well with any amount of missing data and with any type of variable (for example, continuous or dichotomous) (Allison, 2002; Little & Rubin). Rounding is used to assign values to categorical responses. Unbiased slopes and standard errors are produced, and it builds in the normal variability that would be present in a complete data set (Streiner, 2002). Standard errors are even less biased using the "rolled up" method than when using one implicate data set.

The imputation methods described here vary greatly in the amount of time required to learn and use them. The least sophisticated procedure, listwise deletion, typically takes about a minute to process with a statistical software package. On the other end of the spectrum, using the single implicate method took more than four hours on the first use (Table 1).

Table 1: Comparison of Time to Learn and Use Each Imputation Method

Imputation Method (Software)	Time to Conduct Each Method: First Use	Second Use
Listwise (SAS)	1 minute	1 minute
Mean imputation	60 minutes	20 minutes
Hotdecking	180 minutes	100 minutes
Regression imputation (STATA)	45 minutes	18 minutes
First implicate (NORM)	255 minutes	60 minutes
Multiple implicates (NORM)	150 minutes	80 minutes

METHOD

Data Sets

To compare the results from using each of these data imputation methods, we used variables with missing values from two data sets in statistical analyses. These data sets were chosen for demonstration purposes because they were accessible to us and they were of different sizes. There were some commonalities between the data sets, the research teams, and the methods used to create the data collection instruments, for example. In addition, they had similar data entry and data analysis methods and statistical consultants. These commonalities controlled for different methods of instrument development and data entry methods that can be used. The first data set, Service Use of Depressed Elders, sampled elderly people in a community after acute hospitalization for depression (Morrow-Howell, Proctor, Rubin, Li, & Thompson, 2000; Proctor et al., 2003). At the time this analysis was completed, data had been collected from 169 participants at the time of their hospitalization (Table 2). For purposes of this demonstration, the participant's score on the Geriatric Depression Scale (GDS) (Sheikh & Yesavage, 1986) was used as a dependent variable. The participant's GDS score as recorded by hospital staff at intake was taken from the hospital record. Sixteen percent of these scores were missing because the test was not given by the hospital staff because of staff time constraints, the individual's short length of stay, or the participant's refusal to answer the questions.

Three other measures from the data set were used as independent variables: age, self-reported income, and the participant's score on the Mini-Mental State Examination (MMSE) (Cockrell & Folstein, 1988), a test of cognitive ability. None of the data on respondent age were missing; however, income had 21% missing data and 2% of MMSE scores were missing. Again, the participant's refusal to provide information, time constraints of hospital staff, or the patient's short length of stay account for the gaps of data. All of these reasons were considered to possibly form a pattern in the missing values, thus the data were assumed to be MNAR.

Table 2: Description of Data Sets Used in Analysis

Purpose of Study	Sample	Data Collection	Variables Used in Analysis
Service Use of Depressed Elders $N = 169$			
Examines services use and outcomes for elderly people in the community after acute hospitalization for depression.	All individuals 65 or older hospitalized in geropsychiatric unit for depression in large teaching hospital in Midwest and discharged to community setting. About 87% of those eligible participated.	Surveys completed during interviews before hospital discharge, six weeks and six months after hospitalization	GDS Score: range: 0–30; higher scores indicate greater depression Age in years Income: categorical, range 1 (less than $3,000) to 16 ($50,000 or more) MMSE score: range: 1–30; higher scores indicate greater cognitive ability
Youth Services Project $N = 792$			
Examines mental health and behaviors of adolescents	792 adolescents using educational, child welfare, juvenile justice, or primary care sector services (approximately 200 per sector)	Surveys completed at youths' homes or a private place of their choosing	Grade point average: range 0–4, higher score indicating better grades Prostitution in neighborhood: 0 = none, 1 = some, 2 = a lot Neighbors on welfare: 0 = none, 1 = some, 2 = a lot Drug dealing in school: 0 = none, 1 = some, 2 = a lot Depend on family: 1 = rarely to 5 = all the time

Note: The Service Use of Depressed Elders After Acute Care study was conducted at Washington University in St. Louis (Nancy Morrow-Howell, PhD, and Enola K. Proctor, PhD, co-principal investigators) and funded by the National Institute of Mental Health (NIMH), grant no. MH56208. The Youth Services Project study was conducted at Washington University in St. Louis (Arlene Stiffman, PhD, principal investigator) and funded by the NIMH (MH56425). GDS = Geriatric Depression Scale (Sheikh & Yesavage, 1986); MMSE = Mini-Mental State Examination (Cockrell & Folstein, 1988).

The second data set used for this analysis, the Youth Services Project, included data from 792 adolescents in a study of the mental health and behaviors of youths (Hadley-Ives, Stiffman, Elze, Johnson, & Doré, 2000; Stiffman, Hadley-Ives, Elze, Johnson & Doré, 1999) (Table 2). From this study, the self-reported youth's grade point average (GPA) was used as a dependent variable. Only 1% of the data for this variable was missing. Missing values were due to participants not reporting their GPA. The independent variables included measures of prostitution in the neighborhood, the number of neighbors on welfare, the presence of drug dealing in the school, and the youth's ability to depend on family members for support of school achievement. The independent variables had 1% to 9% of the data missing: prostitution in the neighborhood (2%), neighbors on welfare (9%), drug dealing in school (6%), and support from family (1%). Missing data among these independent variables stemmed from participants refusing to answer an item or leaving the item blank. Given that most of the independent variables were related to sensitive issues for these youths, it was assumed that there were patterns to the missingness of the values, and therefore the missing values were assumed to be MNAR.

RESULTS

Comparison of Descriptive Statistics

The first analysis compared the means and standard errors of each of the variables in both data sets after each imputation method was conducted. The loss of statistical power that is common with listwise deletion is clearly demonstrated by the loss of 58 cases, which was 34% of the sample of Service Use of Depressed Elders and 125 cases (16%) in the Youth Services Project sample (Table 3). Using any of the other methods of data imputation allowed all cases to be used in the analysis.

Also noteworthy is the variation in mean and standard error values across the imputation methods. When variables have few missing values (MMSE

Table 3: Comparison of Means and Standard Errors of the Dependent and Independent Variables in Each Data Set

Data Set and Variables	Multiple Implicate	Listwise	Mean Substitution	Hotdecking	Regression Imputation	Single Implicate
Service Use of Depressed Elders	**(*n* = 169)**	**(*n* = 111)**	**(*n* = 169)**	**(*n* = 169)**	**(*n* = 169)**	**(*n* = 169)**
Age (0% missing)						
M	76.03	76.75	76.03	76.03	76.03	76.03
SE	.521	.674	.521	.521	.521	.521
Income (21% missing)						
M	10.25	10.32	10.24	10.41	10.25	10.18
SE	.290	.329	.240	.259	.242	.273
MMSE (2% missing)						
M	23.06	23.29	23.10	23.05	23.09	23.07
SE	.401	.502	.398	.403	.398	.402
GDS (dependent variable) (16% missing)						
M	11.896	12.25	12.30	12.12	12.25	12.02
SE	.640	.651	.506	.564	.509	.589
Youth Services Project	**(*n* = 792)**	**(*n* = 667)**	**(*n* = 792)**	**(*n* = 792)**	**(*n* = 792)**	**(*n* = 792)**
Prostitution in neighborhood (2% missing)						
M	.48	.49	.48	.48	.48	.48
SE	.027	.029	.026	.026	.026	.026
Neighbors on welfare (9% missing)						
M	1.19	1.19	1.20	1.11	1.19	1.18
SE	.030	.029	.028	.028	.028	.027
Drug dealing in school (6% missing)						
M	.63	.66	.63	.63	.63	.63
SE	.027	.028	.025	.025	.026	.25
Depend on family (1% missing)						
M	4.15	4.17	4.15	4.15	4.15	4.15
SE	.041	.044	.041	.040	.041	.040
GPA (dependent variable) (1% missing)						
M	2.25	2.25	2.24	2.25	2.24	2.25
SE	.033	.035	.033	.033	.034	.033

Note: MMSE = Mini-Mental State Examination (Cockrell & Folstein, 1988); GDS = Geriatric Depression Scale (Sheikh & Yesavage, 1986). GPA = grade point average.

and depend on family, for example), there is very little difference among the means and standard errors across the imputation methods (Table 3). However, when larger percentages of missing values are present (income, GDS score, and neighbors on welfare), greater variations across the imputation methods are apparent. The variation appears to be greater on the Service Use of Depressed Elders data set, which has a smaller sample size than the Youth Services Project data set. Compared with the multiple implicate method, the greatest variation among the other five imputation methods appears when the hotdecking method is used, especially with those variables that have a larger percentage of missing values.

Regression Analyses with Each Imputation Method

The three independent variables in the Elders data set (that is, age, income, MMSE score) were regressed on the dependent variable, GDS score. A number of important results should be noted. First, the loss of statistical power using listwise deletion was again clearly demonstrated by the decrease in sample size from 169 to 111, 34% of the sample (Table 4). This is a substantial portion of the cases in

Table 4: Model Statistics for Dependent Variable: Geriatric Depression Scale from the Service Use of Depressed Elders Study

	Multiple Implicate (n = 169)	Listwise (n =111)	Mean Substitution (n = 169)	Hotdecking (n = 169)	Regression Imputation (n = 169)	Single Implicate (n = 169)
Model F	N/A[a]	2.59	2.11	.72	2.64	1.88
p		.056	.10	.54	.052	.13
R^2	.05	.07	.04	.01	.05	.03
Age						
Slope	.088	–.034	.045	.076	.065	.060
SE	.088	.090	.074	.084	.074	.087
t	1.01	–.38	.61	.92	.88	.69
p	.314	.71	.54	.36	.38	.49
Income						
Slope	–.394	–.384	–.365	–.130	–.398	.353
SE	.174	.186	.162	.169	.161	.167
t	–2.26	–2.06	–2.25	–.77	–2.48	–2.12
p	.02	.04	.03	.44	.01	.04
MMSE						
Slope	–.076	–.201	–.068	–.083	–.072	–.030
SE	.120	.122	.097	.109	.097	.113
t	–.63	–1.65	–.70	–.76	–.74	–.27
p	.53	.10	.49	.45	.46	.79

Note: Higher numbers mean greater depression. MMSE = Mini-Mental State Examination (Cockrell & Folstein, 1988). N/A = not applicable.
[a]F values are not averaged.

this study. With such a large loss of cases, the ability of the remaining cases to represent the entire sample and larger population must be carefully considered. If additional variables were added to the equation, the sample size would likely continue to decrease. All of the other imputation methods were able to retain all of the cases and maximize the statistical power, an especially important consideration with small data sets.

Next, the F values for the regression imputation method and listwise deletion appear to be inflated compared with those for the other methods. This finding is similar to those noted by others who have also found that regression imputation is likely to inflate the F values, increasing the possibility of reporting results that inaccurately suggest significance (Graham & Hofer, 2000).

Note the income variable in Table 4 under the hotdecking method. The slope of this variable appears to be considerably higher compared with the other methods. This example illustrates the general caution that hotdecking can be a highly volatile method to use.

Last, the variation in the t values and their associated p values for the MMSE score variable should be noted. None of these values were statistically significant, but the t values range from –.27 (single implicate) to –1.65 (listwise deletion). Similarly, the p values range from .10 (listwise deletion) to .79 (single implicate), a range of .69. If this degree of variation in p values occurred around the α = .05 significance level, a researcher would be reporting statistical significance where it may not exist or not reporting significance where it may very well exist, depending on the imputation method he or she used.

We conducted the same regression analyses on the Youth Services Project data set. For this regression equation, the independent variables (prostitution in the neighborhood, welfare recipients in the neighborhood, presence of drugs in school, participant's ability to depend on family members) were regressed on each participant's GPA. As noted with the first data set, there was a significant loss of cases using the listwise deletion method: 125 cases, or 16% of the 792 respondents (Table 5). This loss of cases may not significantly reduce the statistical power for analyses with this data set given the larger sample size. However, a 16% loss of the cases is a large enough portion of the sample to question

Table 5: Model Statistics for Dependent Variable: Grade Point Average from the Youth Services Project Study	Multiple Implicate (n = 792)	Listwise (n = 667)	Mean Substitution (n = 792)	Hotdecking (n = 792)	Regression Imputation (n = 792)	Single Implicate (n = 792)
Model F		6.10	6.48	6.70	6.79	6.37
p	N/A[a]	<.001	<.001	<.001	<.001	<.001
R^2		.036	.032	.033	.033	.031
Prostitution in neighborhood						
Slope	–.080	–.107	–.086	–.082	–.084	–.088
SE	.049	.050	.047	.047	.047	.047
t	–1.65	–2.14	–1.83	–1.75	–1.78	–1.86
p	.099	.032	.068	.080	.0746	.063
Neighbors on welfare						
Slope	–.029	–.011	–.022	–.012	–.028	–.004
SE	.050	.048	.043	.044	–.046	.045
t	.57	–.23	–.50	–.27	–.62	–.08
p	.569	.815	.616	.790	.534	.934
Drugs in school						
Slope	–.153	–.152	–.160	–.171	–.162	–.154
SE	.049	.050	.048	.047	.047	.047
t	–3.13	–3.06	–3.33	–3.65	–3.39	–3.24
p	.0019	.0020	.0009	.0003	.0007	.0012
Depend on family						
Slope	.066	.066	.062	.059	.0637	.068
SE	.030	.031	.029	.029	.0289	.029
t	2.24	2.10	2.16	2.03	2.20	2.33
p	.025	.036	.031	.042	.028	.020

Note: Grade point averages were from the last semester before the participants' interview.
[a]F values are not averaged.

whether the remaining cases are representative of the entire group.

Several other observations are also noteworthy. First, the greatest variation in the slopes across imputation methods occurs with the variable neighbors on welfare, which had the largest amount of missing values (9%) in this data set. This suggests that even large samples cannot "correct" for all variations in the sample. In fact, some of the findings with this larger data set with smaller amounts of missing data are similar to those found in the smaller Service Use of Depressed Elders data set. For example, the p values for the prostitution in the neighborhood variable vary from the .032 level (statistically significant) to the .099 level (not statistically significant). As discussed earlier, the variation in the p values across the data imputation methods suggests that researchers would or would not report statistical significance depending on the method used. Also worth mentioning is that listwise deletion, typically used by social work researchers, is the only method that suggests statistical significance of this variable.

Prostitution in the neighborhood is an example of a variable that probably should be considered MNAR because of the sensitive nature of the item. Some teenagers may be uncertain of what constitutes prostitution, making their responses unreliable; others may be embarrassed to admit that prostitution occurs in their neighborhood and respond negatively. As noted earlier, researchers need to consider the nature of each item in their data to identify any possible patterns in missing values.

DISCUSSION

Missing data, often ignored during data analysis, contribute to biased results, making it difficult to make valid and efficient inferences about a population to guide both practitioners and researchers (Schafer & Graham, 2002). Researchers are most likely to ignore missing data when they fail to understand the significance of the problem or lack an understanding of possible solutions to address the problem (Figueredo et al., 2000; Orme & Reis, 1991). This analysis exposed the hazards of ignoring missing data and examined six data imputation methods to address this problem: listwise deletion, mean substitution, hotdecking, regression imputation, single implicate, and multiple implicate. As virtually all social science research will inevitably have some incomplete data, missing data should be a universal concern and addressed in all research reports. Social workers must join their colleagues in related disciplines who more consistently conduct and report this aspect of their research studies.

The results of the statistical analysis conducted for this study suggest that a large sample with only a small percentage of missing values is not influenced to the same degree by data imputation methods as are smaller data sets. However, regardless of the sample size, researchers should still consider the advantages and disadvantages in choosing the most appropriate imputation method. As with all decisions related to a research study, those related to data imputation methods must be informed and made within professional guidelines for the conduct of research. For example, a researcher with a data set with only a small amount of missing data must decide whether the time required to use the multiple implicate method, statistically the gold standard but also the most time-intensive method, is warranted or whether mean substitution or listwise deletion, very quick methods, could be used equally as well. In this study, there were few differences found between these methods when used with variables missing just 1% or 2% of the values.

The researcher must also carefully consider the amount of missing values and whether the variable is to be used as an independent or dependent variable. Researchers are likely to want to retain a dependent variable and therefore may be more inclined to accept higher levels of missing values. Although there does not appear to be a clear definition of how much data can be imputed, the literature suggests that 20% or less is acceptable (Little & Rubin, 2002). When missing values exceed these guidelines, the best decision may be to identify an alternative variable for use in the analysis. In all cases, researchers should clearly document the amount of missing data and their decisions regarding imputation methods in any research reports.

Even after being convinced that missing data bias results and should be addressed, the perception that these methods require too much time may make some researchers hesitant to use them. The results from this study suggest that decisions about data imputation methods should not be made strictly on the amount of time necessary to use them. For example, listwise deletion required the least amount of time but the results were more biased, which could lead the researcher to incorrectly report significance levels. The most time-consuming methods in this study were the single implicate and multiple implicate methods; however, these also produced the most accurate values. Given the number of hours spent overall on a research project, the additional time necessary to learn and use these methods is well worth the investment.

The final verdict on multiple imputation methods is not being handed down in this article. Although beyond the scope of this article, conducting Monte Carlo studies with the two data sets used in this analysis would significantly add to our understanding about imputation methods. Much of the information that is available on imputation methods has been developed theoretically and tested by statisticians but not published in academic journals or textbooks. In sum, we need a more accessible literature on multiple imputation methods.

In conclusion, everyone recognizes that the easiest way to handle missing data is to avoid missing values during the data collection process. However, given the type of research that social workers typically conduct, avoiding missing values altogether is not realistic. There is a wide variability in knowledge, skills, and resources available to social work researchers, and all of these factors influence the quality of the research analysis possible and subsequently reported in the literature.

Given these realities, our review of the literature, and the results of this study, the following recommendations are made:

- Every researcher should explore the patterns of missing values in data set and consider constructing instruments to clearly identify some patterns of missingness (for example, "refuse" or "don't know").
- Social work can no longer avoid the issues of missing data. Every research report should

report the reasons for and the amount of missing data as well as what data imputation method was used during the analysis.

- Multiple implicate is currently the best imputation method and should be used whenever possible. Other methods should only be considered when working with a large data set that has values MCAR.
- If resources (time, skills, money) prevent the researcher from using the multiple implicate method, the researcher should understand the implications of biased estimates and how other researchers may produce different results when trying to replicate the study.
- In reporting, researchers must make methods transparent so that another researcher could reproduce the analysis and get the same results.

The last two recommendations address the very foundation of the scientific process—that is, being able to reproduce a research study to substantiate or refute it. Biased estimates, which are more likely to result from using data imputation methods other than multiple implicate, severely threaten the reproducibility of research. By using the most sophisticated methods available, the quality of data analysis and associated reports of that research will be enhanced. This can only strengthen the profession's knowledge base for practice, research, and theory development. SWR

REFERENCES

Allison, P. D. (2002). *Missing data* (Sage University Paper Series on Quantitative Applications in the Social Sciences 07–136). Thousand Oaks, CA: Sage Publications.

Cockrell, J. R., & Folstein, M. F. (1988). Mini-Mental State Examination (MMSE). *Psychopharmacy Bulletin, 24*, 689–692.

Figueredo, A. J., McKnight, P. E., McKnight, K. M., & Sidani, S. (2000). Multivariate modeling of missing data within and across assessment waves. *Addiction, 95*(Suppl. 3), S361–S380.

Graham, J. W., & Hofer, S. M. (1993). *EMCOV reference manual.* Los Angeles: University of Southern California, Institute for Prevention Research.

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 multilevel data: Practical issues, applied approaches, and specific examples* (pp. 201–218). Mahwah, NJ: Lawrence Erlbaum.

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

Hadley-Ives, E., Stiffman, A. R., Elze, D., Johnson, S. D., & Doré, P. (2000). Measuring neighborhood and school environments: Perceptual and aggregate approaches. *Journal of Human Behavior in the Social Environment, 3*(1), 1–28.

Little, R.J.A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: John Wiley & Sons.

McDonald, R. A., Thurston, P. W., & Nelson, M. R. (2000). A Monte Carlo study of missing item methods. *Organizational Research Methods, 3*, 71–92.

Morrow-Howell, N. L., Proctor, E. K., Rubin, E. H., Li, H., & Thompson, S. (2000). Service needs of depressed older adults following acute psychiatric care. *Aging & Mental Health, 4*, 330–338.

Orme, J. G., & Reis, J. (1991). Multiple regression with missing data. *Journal of Social Service Research, 15*(1/2), 61–91.

Pigott, T. D. (2001). A review of methods for missing data. *Educational Research and Evaluation, 7*, 353–383.

Proctor, E. K., Morrow-Howell, N. L., Doré, P., Wentz, J., Rubin, E. H., Thompson, S., & Li, H. (2003). Comorbid medical conditions among depressed elderly patients discharged home after acute psychiatric care. *American Journal of Geriatric Psychiatry, 11*, 329–338.

Raaijmakers, Q.A.W. (1999). Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement, 59*, 725–748.

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

Rubin, D. B. (1991). EM and beyond. *Psychometrika, 56*, 241–254.

SAS Institute Inc. (2000). *SAS language and procedures.* Cary, NC: Author.

Schafer, J. L. (1998). *NORM: Multiple imputation of multivariate continuous data under a normal mode. Version 2.* [Computer software for Windows 95/98/NT]. Available at http://www.stat.psu.edu/~jls/misoftwa.html

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

Sheikh, J. I., & Yesavage, J. A. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. In T. L. Brink (Ed.), *Clinical gerontology: A guide to assessment and intervention* (pp. 165–173). New York: Haworth Press.

SPSS, Inc. (2002). *SPSS for Windows 10.0.* Chicago: Author.

Stata. (2001). *Stata user's guide.* College Station: TX: Author.

Stiffman, A.R., Hadley-Ives, E., Elze, D., Johnson, S., & Doré, P. (1999). Impact of environment on adolescent mental health and behavior: Structural equation modeling. *American Journal of Orthopsychiatry, 69*, 73–86.

Streiner, D. L. (2002). The case of the missing data: Methods of dealing with dropouts and other research vagaries. *Canadian Journal of Psychiatry, 47*, 68–75.

Tabachnick, B. G., & Fidell, L. S. (1983). Cleaning up your act: Screening data prior to analysis. In B. G. Tabachnick & L. S. Fidell, *Using multivariate statistics* (pp 68–81). New York: Harper & Row.

***Jeanne A. Saunders, PhD,** is assistant professor, School of Social Work, University of Iowa, 308 North Hall, Iowa*

City, IA 52242; e-mail: jeanne-saunders@uiowa.edu. **Nancy Morrow-Howell, PhD**, is Ralph and Muriel Pumphrey Professor of Social Work, George Warren Brown School of Social Work; **Edward Spitznagel, PhD**, is professor, Department of Mathematics; **Peter Doré, MA**, is systems manager, Center for Data Management, George Warren Brown School of Social Work; and **Enola K. Proctor, PhD**, is Frank J. Bruno Professor of Social Work Research and dean for research, George Warren Brown School of Social Work, Washington University, St. Louis, MO. **Richard Pescarino, MS**, is an instructor, St. Louis Community College, Florissant Valley, St. Louis, MO. Address all correspondence to Dr. Saunders.

Original manuscript received December 8, 2003
Final revision received September 28, 2004
Accepted June 7, 2005