Journal of Modern Applied Statistical Methods

Volume 13 | Issue 2 Article 3

11-2014

Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research

Jehanzeb R. Cheema
University of Illinois at Urbana-Champaign, jcheema1@masonlive.gmu.edu

Journal of Modern Applied Statistical Methods
November 2014, Vol. 13, No. 2, 53-75.

ISSN 1538 − 9472

Regular Articles: Some General Guidelines for Choosing Missing Data Handling Methods in Educational Research

Jehanzeb R. Cheema
University of Illinois at Urbana-Champaign
Champaign, IL

The effect of a number of factors, such as the choice of analytical method, the handling method for missing data, sample size, and proportion of missing data, were examined to evaluate the effect of missing data treatment on accuracy of estimation. A methodological approach involving simulated data was adopted. One outcome of the statistical analyses undertaken in this study is the formulation of easy-to-implement guidelines for educational researchers that allows one to choose one of the following factors when all others are given: sample size, proportion of missing data in the sample, method of analysis, and missing data handling method.

Keywords: Missing data, imputation, simulation, listwise deletion, missing value analysis

Introduction

Missing data is an issue that most researchers in education encounter on a routine basis. In survey research there can be many reasons for missing data such as respondents ignoring a few or all questions, questions being irrelevant to the respondent's situation, or inability of survey administrators to locate the respondent. Missing data can also occur in non-survey data, such as experimental and administrative data (Acock, 2005; Brick & Kalton, 1996; Groves et al., 2004). In non-survey samples, missing data can arise due to carelessness in observation, errors made during data entry, data loss due to misplacement etc. Regardless of the reason why data is missing, once it is missing it becomes part of the dataset that is then used by researchers to perform analytical procedures. The quality of such

Dr. Cheema is a Clinical Assistant Professor in the Bureau of Educational Research, College of Education. Email him at jrcheema@illinois.edu.

analytical procedures directly depends on the quality of underlying data which in turn can be affected by the nature of missing data (Allison, 2001; Schafer & Graham, 2002).

Unfortunately there are many different methods of handling missing data which can have profoundly different effects on estimation. For this reason it is important to select the correct missing data handling method that is suited to a researcher's particular circumstances. These circumstances can be expressed as factors, such as sample size, proportion of missing data, method of analysis etc., some of which may fall under the control of the researcher in a given scenario and thus can be manipulated, while others are more difficult to control.

For example, a researcher working with secondary data will likely not find it possible to increase the sample size to offset the effect of missing data but may have flexibility regarding the choice of analytical method. On the other hand, a researcher who is gathering her own data and who is relying on a specific method of analysis to answer her research questions may find it easy to increase her sample size in order to lower the proportion of missing cases. As these illustrations suggest, the scenario under which a researcher handles missing data can vary considerably depending on that researcher's circumstances.

There were many investigations and comparisons of the performance of missing data handling methods, both in general (Afifi & Elashoff, 1966; Graham, Hofer, MacKinnon, 1996; Haitovsky, 1968; Peng, Harwell, Liou, & Ehman, 2009; Peugh & Enders, 2004; Wayman, 2003; Young, Weckman, & Holland, 2011) and in context of specific factors such as proportion of missing data (Alosh, 2009; Knol et al., 2010; Rubin, 1987) and sample size (Alosh, 2009; Rubin, 1987). Because the current study is not a review of the literature, any comprehensive attempt to reproduce that discussion is beyond its immediate scope. For detailed technical aspects including mathematically-intensive proofs and theorems, and application of these methods in various fields including education, see Madow, Nisselson and Olkin (1983), Madow and Olkin (1983), Madow, Olkin, and Rubin (1983), Jones (1996), Groves, Dillman, Eltinge, and Little (2002), and Andridge & Little (2010).

Although several researchers have investigated missing data handling methods, their results were based on various combinations of sample size, proportion of missing data, method of analysis, and missing data handling method. None of the past studies has dealt with all of these factors simultaneously using the same dataset in order to control for data-specific characteristics. For this reason, the findings of these earlier studies cannot be used to construct general guidelines for use with new datasets. This study controls for all of these factors simultaneously, and also expands the range of sample size and proportion of missing data in order

to improve the generalizability of its findings. Furthermore, in this study the missing data handling methods are compared for four analytical methods that are frequently employed in educational research: one sample *t* test, independent samples *t* test, two-way ANOVA, and linear multiple regression. Results of these comparisons can be used to correct biases in tests of hypotheses reported in past research that employed improper imputation methods, such as mean imputation, that are well-known to produce biased parameter estimates.

Even though the drawbacks of many missing data handling methods are well-known and have been regularly publicized in leading peer-reviewed journals, researchers in social sciences in general and education and psychology in particular have shown a remarkable resilience in sticking to some of the simpler and most error-prone methods such as listwise deletion, pairwise deletion, and mean imputation (Peng et al., 2006; Peugh & Enders, 2004; Roth, 1994; Schafer & Graham, 2002). There are various reasons for avoiding sophisticated missing data handling methods that range from a lack of expertise in quantitative methodology required for a basic understanding of these methods to the inability to practically implement those methods using specialized software programs due to a lack of programming know-how. A correction of this state of affairs requires a study that specifically targets this population of researchers and that can provide general guidelines for selection of the best missing data handling method under a variety of scenarios. Some prior studies such as Roth (1994) have pointed out the absence of an expansive measurement of bias due to missing data and the gain in efficiency that can be achieved by imputing that data in social science literature, especially psychology, a field from which educational research heavily borrows its quantitative methodology. The same study especially stressed development of guidelines that can be used to choose the best missing data handling technique in a variety of circumstances faced by researchers.

The main objective of this study is to provide educational researchers with general guidelines about which missing data handling method performs best under a variety of combinations of sample size, proportion of missing data, and method of analysis. More specifically, these guidelines will allow the researcher to choose one of the following factors when all others are given: sample size, proportion of missing data, method of analysis, and missing data imputation method.

Method

The analytical procedures presented in this study use two sources of data, a simulated dataset and empirical samples. A description of these datasets and analytical procedures follows.

Data Simulation

The primary source of data used for statistical analyses performed in this study was a simulated dataset. The main reason for using simulated data was to ensure that distributional assumptions governing the methods of analysis applied in this study were not violated. The main concern was that violation of underlying model assumptions for each method of analysis under some conditions and not the others can significantly erode uniformity of the basis on which these methods are compared. A reliable way to avoid this problem was to simulate data that satisfied all underlying assumptions for analytical methods of interest and that at the same time had characteristics that made such data suitable for analysis of real-world problems.

In order to mimic data routinely encountered by educational researchers a dataset with 10,000 cases was simulated which included four continuous and one categorical variable. Because groups of variables are usually investigated because they are related to each other, it is important that the simulated data also mimic such relationships. This was achieved by specifying a variance-covariance matrix that was not unlike what a typical educational researcher may encounter during her research.

The four continuous variables, Y, X_1, X_2, and X_3, were generated in such a way as to simulate weak correlation between Y and X_1 (r = .3), moderate correlation between Y and X_2 (r = .5), and strong correlation between Y and X_3 (r = .7), with the three X's correlated weakly with each other (r = .2). This pattern was adopted to avoid the problem of multicollinearity in linear multiple regression models analyzed in this study. It should be noted that the strength of an association is a relative concept. While a coefficient of correlation of .7 may be considered weak in context of a physical experiment, the same might be considered very strong in context of a social study. Cohen (1992), for instance, suggests .1, .3, and .5 as rule of the thumb for small, medium, and strong correlation. Values of the four continuous variables X_1, X_2, X_3, and Y were drawn from a multivariate normal distribution. For ease of interpretation all continuous variables were specified to have a mean of 0 and standard deviation of 1. Dichotomous predictor Z_1 was

constructed using a uniform discrete distribution with values 0 (n = 4,945) and 1 (n = 5,055).

Because the assignment of these values to Z_1 is random, this mirrors a situation where a significant mean difference in Y does not exist across levels of Z_1. In order to construct the opposite scenario where mean differences do exist, Z_2 was constructed to have three levels, with mean Y significantly different between these levels. The three levels of Z_2 were labeled 1 (n = 1,623), 2 (n = 6,823), and 3 (n = 1,554) with mean Y being the largest for group 1 and smallest for group 3. It should be noted that even though this means that the pattern of missing data in Y now depends on Z_2, such dependency rules out only the missing completely at random (MCAR) assumption and not the relatively less stringent missing at random (MAR) assumption and as the missing values of Y are still independent of their own magnitude, the data cannot be considered as not missing at random (NMAR).

Data Analysis Approach for Simulated Data

The simulated dataset (n = 10,000) was used to select 10 sub-samples of size 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000. Each of these sub-samples was then reduced in size by 1%, 2%, 5%, 10%, and 20% in order to simulate datasets containing missing data. The cases were discarded randomly from each complete sample five times separately in order to make sure that there were no dependencies between samples. Each of the five missing data handling methods were applied to all samples containing missing data under four methods of analysis. These methods of analysis are one sample t test, independent samples t test, two-way ANOVA, and multiple regression.

The main considerations behind the choice of these four methods of analysis is their widespread use among educational researchers and the desire not to restrict the findings of this study to a single method of analysis. These methods represent various modeling regimes encountered routinely by researchers in education. For the independent samples t test, the mean difference in Y over levels of Z_1, the only categorical predictor with two levels, was analyzed. For two-way ANOVA, both categorical predictors, Z_1 and Z_2 were used as factors of Y. And for multiple regression, Y was specified as a function of the three X's and Z_1. Five missing data handling methods were selected for missing data analysis. These methods are listwise deletion, mean imputation, regression imputation, maximum likelihood imputation (ML), and multiple imputation. These methods were chosen because of their ready availability and easy implementation in general statistics packages such as SPSS. Application of these five missing data handling methods under various

sample sizes with data missing in various proportions, and for different methods of analysis forms the core of simulated data analysis.

For each of the four methods of analysis, model parameter estimates and associated tests of hypotheses were obtained generated for the 10 complete and 50 partial samples using each of the five missing data handling methods. In other words, a total of $4 \times (10 + 50) \times 5 = 1{,}200$ models were fitted. These 1,200 models can be categorized into two groups with the first group comprising of 200 models based on samples that contain no missing data and the second group comprising of 1,000 models based on samples that contain missing data. The model significance for these two groups was then compared using the *t* statistic for models involving one sample *t* test and independent samples *t* tests, and the *F* statistic for two-way ANOVA and multiple regression models.

For example, the *F* statistic evaluating model significance for two-way ANOVA under multiple imputation of missing data when the sample size is 100 and proportion of missing data is 5% can be directly compared with the corresponding *F* statistic for the complete sample containing no missing data ($n = 100$). Such a comparison is fair because after imputation the numerator and denominator degrees of freedom are the same for both *F* values. Thus, since the two samples are identical in all other respects including power, any fluctuation in the observed value of *F* can be attributed to the deviation of imputed values from their true counterparts. Such an approach allows an objective evaluation of the effect of an imputation method on the statistic used to test for model significance. For instance if the observed *F* value increases after imputation of missing data, it means that the observed probability of making a Type I error, i.e. rejecting H_0 when H_0 should not be rejected, has decreased.

In order to compare performance of the 1,000 models based on missing data with their complete-data counterparts, a unitless standardized measure of error, the normalized root mean squared error (*RMSE*) was utilized. *RMSE* is in essence the average distance of observed error from the true value and can be interpreted as the standard deviation of $X_{Observed}$. This measure thus takes into consideration the absolute size of error. However, *RMSE* calculated in this way has the same unit of measurement as *X*. By dividing *RMSE* with the range of *X*, the unit of measurement can be removed from *RMSE*. The resulting statistic is called the normalized *RMSE*. The advantage of using normalized *RMSE* over *RMSE* is that it can be used to compare error across variables that are not based on the same unit of measurement.

Empirical Sample 1

In order to test the real-world applicability of simulated results, a large scale dataset with variables having characteristics similar to those used in the simulated data was utilized. This empirical data was obtained from U.S. portion of the Program for International Student Assessment (PISA) (NCES, 2003) which is an assessment of literacy in mathematics, reading, and science of 15-year old students (n = ,456). The questionnaire for this survey was the basis for a large number of variables, some of which are comparable to those simulated in this study. The primary idea behind using an empirical sample was to test the effectiveness of guidelines constructed on the basis of simulated data. The variable selection was based on similarity of characteristics of these variables with their simulated counterparts.

The dependent variable was math achievement which was distributed normally, measured on a continuous scale, and ranged between 200 and 800. Three continuous variables were chosen as predictors of math achievement on the basis of similarity between the variance-covariance matrix of these predictors and that of the simulated continuous variables. These predictors are reading achievement, math anxiety, and the index of home educational resources. Reading achievement was normally distributed and ranged between 200 and 800. Math anxiety is a measure of anxiety felt by a student when engaged in math-related tasks. This variable was measured on a continuum, was normally distributed, and standardized to have a mean of 0 and standard deviation of 1.

Home educational resources measured educational resources owned by a student's household and can be roughly thought of as a component of the student's socioeconomic status. The variable was also standardized to have a mean of 0 and standard deviation of 1. A comparison between the variance-covariance matrices of simulated and empirical predictors showed slight differences in magnitude. However, what is more important to note is the similarity in the pattern of relationship among the four variables which showed that math achievement was correlated somewhat weakly with home educational resources (r = .3), moderately with math anxiety (r = −.4), and strongly with reading achievement (r = .8). This pattern was not very different from that simulated for Y and its three continuous predictors. Similarly, the inter-predictor correlations presented were also weak like their simulated counterparts ranging between −.3 and .3.

The observed deviation between these two variance-covariance structures emphasizes the practical difficulty associated with obtaining empirical datasets which possess exact distributional characteristics that a researcher may require. In addition to continuous variables a categorical predictor, gender was selected from the PISA 2003 dataset. Gender has two categories: male, n = 2,740; and female,

n = 2,715. One case had a missing value for gender reducing the maximum number of observations available for analysis from 5,456 to 5,455.

For the purposes of this study, the approach used for simulated dataset was replicated with PISA data. This allows us to compare estimation results with and without missing data imputation. For analysis, we predicted math achievement from its predictors using a linear multiple regression equation.

The empirical variables were used to evaluate the effectiveness of missing data handling guidelines formed with simulated data. A portion of the empirical dataset was designated as missing and was then analyzed using the same missing data handling methods that were employed for simulated data analysis. This involved selecting an appropriate analytical method, estimating model parameters, and then comparing the estimation results for complete dataset with its incomplete and imputed counterparts in order to evaluate whether the differential effects of missing data handling methods.

Empirical Sample 2

A smaller empirical dataset was employed in order to evaluate the effectiveness of missing data handling methods for small datasets. This data comes from the Population and Housing portion of decennial U.S. Census published by the U.S. Census Bureau (2000). The data chosen for this example is for the states of Virginia and Wisconsin and includes the percentage of individuals in each county with at least a four year college degree for the year 2000. The dataset consists of 207 counties (Virginia, *n* = 135; Wisconsin, *n* = 72).

As with empirical sample 1, the objective of using this sample was to illustrate the effect of missing data handling methods on accuracy of estimation. This was accomplished by specifying a portion of the data as missing using a subset of the missing data percentages used for the simulated dataset. Next, missing data were imputed and the parameter estimates obtained with and without imputation were compared in order to evaluate the effect of various missing data handling methods. In contrast to empirical sample 1 for which a relatively advanced method of analysis viz. multiple regression was employed, for empirical sample 2 a simpler method viz. independent samples *t* test was used to ensure a broader coverage of analytical methods chosen for this study.

Results

Simulated Data

Results of analytical procedures described in the method section for the simulated dataset are presented in this section. In order to see the association between original and imputed data, Pearson coefficient of correlation was calculated between original data and imputed data separately for each imputation method. These correlations were significantly different from zero at 5% level of significance and showed a general decreasing trend in magnitude as the percentage of missing data increased.

Furthermore, the correlations tended to be stronger for maximum likelihood (ML) imputation and multiple imputation methods as compared to mean imputation and regression imputation. When proportion of missing data was 5% or less, almost without exception, all imputation methods produced correlations between original and imputed data that were in excess of .95. Only for sample sizes that were less than 50 with percentage of missing data exceeding 5% did we observe somewhat weaker correlations, in one case falling as low as .74. Mean imputation seemed to work well as long as the percentage of missing data was 10% or less but the correlation between mean imputed and original data fell quickly regardless of sample size as this percentage exceeded 10%. The mean correlation (i.e. correlations averaged over sample size and percentage of missing data) between original and imputed data for mean imputation, regression imputation, maximum likelihood imputation, and multiple imputation were .95, .96, .98, and .98 respectively, suggesting that such correlation was strongest for ML and multiple imputation methods and weakest for mean imputation. However, it should be noted that the difference in magnitude of these correlations is very small.

An examination of normalized *RMSE* values (see Figure 1) showed that multiple imputation was the best missing data handling method because it produced the smallest normalized *RMSE* for all four methods of analysis, one sample *t* test, independent samples *t* test, two-way ANOVA, and multiple regression. For one sample *t* test, all imputation methods performed better than listwise deletion although the difference between listwise deletion and mean imputation was small. For independent samples *t* test, listwise deletion did not perform very well but mean imputation did. Furthermore, for independent samples *t* test, the performance of mean imputation and ML imputation was almost the same. For two-way ANOVA, listwise deletion was as good as ML imputation and better than regression imputation and mean imputation, the latter being the most error-prone method. For

multiple regression, regression imputation worked almost as well as multiple imputation which produced the smallest normalized *RMSE*, listwise deletion and ML imputation behaved similarly, and mean imputation was clearly inferior to all other missing data handling methods.

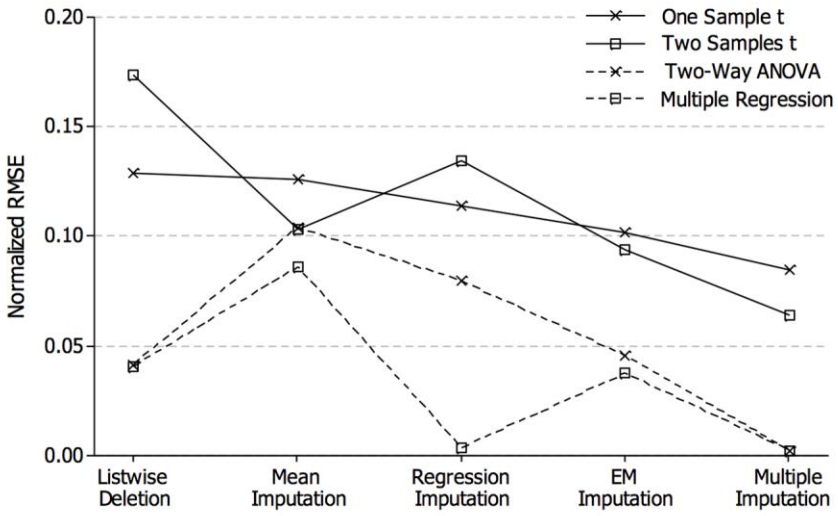

Figure 1. The average effect of missing data handling method on accuracy of estimation for various methods of analysis

The reason why regression imputation performed so well when the analytical method was multiple regression was that using regression-imputed data in a regression equation, when the variables used for imputation and model estimation are the same, is akin to fitting a regression equation twice to predict the same dependent variable. It is important to note here that the results presented in Figure 1 were averaged over sample size and proportion of missing data and therefore cannot be used to evaluate the partial effect of these two factors.

In fact, such averaging contributes to observance of some contradictory results. For example, we see in Figure 1 that mean imputation does not work very well in case of one sample *t* test but does work well for independent samples *t* test even though both methods involve a similar kind of dependence on the sample mean of *Y* and its standard error. For this reason, it is essential that we disaggregate the results in order to clarify the partial effects of sample size and proportion of missing data.

Disaggregated results showed that for one sample *t* test: with small samples ($n \leq 50$), ML imputation worked best whether the proportion of missing data was low ($m \leq 5\%$) or high ($m > 5\%$.); with medium samples ($50 < n < 1{,}000$), multiple imputation worked best regardless of proportion of missing data; and with large samples ($n \geq 1{,}000$), ML imputation works best when proportion of missing data was low and multiple imputation worked best when proportion of missing data was high. It should be noted here that even though we have identified the best missing data method under various conditions, in practical terms the increase in efficiency gained due to applications of that best method may be too small to justify such application.

Power comparisons for the four methods of analysis suggested that with listwise deletion and medium effect sizes as defined by Cohen (1992): one sample *t* test achieved a power of .8 at sample sizes between 20 and 50 for any proportion of missing data ranging between 1% and 20%; independent samples *t* test achieved a power of .8 at sample sizes between 100 and 200 for any proportion of missing data ranging between 1% and 20%; 2×3 ANOVA achieved a power of .8 at sample sizes between 200 and 500 for any proportion of missing data ranging between 1% and 20%; and multiple linear regression with one set of four predictors achieved a power of .8 at sample sizes between 50 and 100 for any proportion of missing data ranging between 1% and 10% and, at sample sizes between 100 and 200 when the proportion of missing data was 20%. It should be noted that for the four imputation methods, power values at all sample sizes were exactly identical to those of the complete data because after imputation sample sizes are at their maximum.

Statistical results for efficiency gains are summarized as a decision tree in Table 1. Out of the 24 possible situations listed in Table 1 based on various combinations of method of analysis (one sample *t* test, independent samples *t* test, two-way ANOVA, multiple regression), sample size (small, medium, large), and proportion of missing data (low, high), relative to listwise deletion, in 15 cases (62.5%) the best method was multiple imputation, in seven cases (29.1%) the best method was maximum likelihood imputation, in only one case (4.2%) the best method was regression imputation, and in only one case (4.2%) the best method was mean imputation. However, the increase in efficiency gained in each of these 24 cases was not the same. For example when multiple regression is the method of analysis, sample size is small, and proportion of missing data is high, the gain in accuracy, defined as the reduction in normalized root mean squared error between the most efficient missing data handling method (multiple imputation in this scenario) and listwise deletion is only about 1%. Thus, in terms of the time and effort required for application of multiple imputation of missing data a researcher

may not find it worthwhile to implement missing data imputation at all rather relying on listwise deletion and be content with the corresponding 1% loss in accuracy that could have been gained otherwise.

Table 1. Summary of gain in estimation accuracy from application of missing data handling methods for various methods of analysis

	One Sample *t* Test						**Independent Samples *t* Test**					
Sample size[a]	*Small*		*Medium*		*Large*		*Small*		*Medium*		*Large*	
Incidence of missing data[b]	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Most efficient data handling method[c]	ML	ML	MI	MI	ML	MI	ML	MI	R	MI	MI	MI
Gain in accuracy[d]	0.07	0.01	0.05	0.11	0.01	0.07	0.06	0.01	0.07	0.04	0.15	0.24

	Two-Way ANOVA						**Multiple Regression**					
Sample size	*Small*		*Medium*		*Large*		*Small*		*Medium*		*Large*	
Incidence of missing data	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High
Most efficient handling method	MI	MI	MI	MI	MI	MI	ML	MI	ML	MI	ML	MI
Gain in accuracy from imputation	0.04	0.03	0.03	0.09	0.02	0.10	0.04	0.01	0.03	0.12	0.02	0.10

Note. EM = Expectation maximization imputation. M = mean imputation. MI = multiple imputation. R = Regression imputation; [a]Small, $n \leq 50$; Medium, $50 < n < 1{,}000$; Large, $n \geq 1{,}000$; [b]Low, missing $m \leq 5\%$; High, missing $m > 5\%$; [c]Most efficient data handling method is the one that produces smallest normalized root mean squared error; [d]Gain in accuracy is measured as the reduction in normalized root mean squared error between the most efficient missing data handling method and listwise deletion. When multiplied by 100 this gain can be interpreted as a percentage.

Empirical Sample 1

In order to allow comparison with simulated data results, a multiple regression equation was used to predict math achievement from reading achievement, math anxiety, home educational resources, and gender. Results for the full dataset and the datasets based on various missing data handling methods are presented in Tables 2 and 3 under low, $m = 5\%$ ($n = 5{,}182$) and high, $m = 10\%$ ($n = 4{,}910$) missing data conditions respectively.

Table 2. Predicting math achievement: multiple regression results with 5% missing data under various missing data handling methods using PISA 2003 data

	Partial Slope Coefficient Estimates					
Predictors	Full Data	Listwise Deletion	Mean Imputation	Regression Imputation	EM Imputation	Multiple Imputation
Intercept	47.20***	48.12***	72.40***	48.39***	48.12***	48.29***
Gender[a]	30.57***	30.26***	28.31***	30.10***	30.23***	30.18***
Home educational resources	0.68	0.79	0.48	0.76	0.79	0.79
Math anxiety	-11.48***	-11.38***	-11.16***	-11.47***	-11.38***	-11.43***
Reading achievement	0.85***	0.85***	0.80***	0.84***	0.85***	0.84***
Model summary						
F	7676.57***	7229.55***	5494.187***	7640.99***	8056.30***	7669.93***
R^2	.849***	.848***	.801***	.849***	.855***	.849***
Power	1.000	1.000	1.000	1.000	1.000	1.000

Note. n = 5,455. F = Observed F from regression ANOVA. R^2 = proportion of explained variance; [a]Reference category is female; * $p < .05$; ** $p < .01$; *** $p < .001$

Table 3. Predicting math achievement: multiple regression results with 10% missing data under various missing data handling methods using PISA 2003 data

	Partial Slope Coefficient Estimates					
Predictors	Full Data	Listwise Deletion	Mean Imputation	Regression Imputation	EM Imputation	Multiple Imputation
Intercept	47.20***	48.36***	88.21***	47.24***	48.36***	48.84***
Gender[a]	30.57***	30.35***	27.53***	30.65***	30.35***	30.22***
Home educational resources	0.68	0.82	0.58	0.89*	0.82	0.86
Math anxiety	-11.48***	-11.72***	-11.04***	-11.73***	-11.72***	-11.83***
Reading achievement	0.85***	0.84***	0.77***	0.85***	0.84***	0.84***
Model summary						
F	7676.57***	6927.02***	4633.834***	7667.110***	8462.10***	7638.26***
R^2	.849***	.850***	.773***	.849***	.861***	.849***
Power	1.000	1.000	1.000	1.000	1.000	1.000

Note. n = 5,455. F = Observed F from regression ANOVA. R^2 = proportion of explained variance; [a]Reference category is female; * $p < .05$; ** $p < .01$; *** $p < .001$

These results show that with the exception of mean imputation, all missing data handling methods produce regression parameter estimates and model statistics such as R^2 and overall F for regression ANOVA that are very similar to their full data counterparts. Almost without exception, the results of tests of hypothesis from each of the models presented in Tables 2 and 3 are identical. The only exception is when regression imputation is used under the 10% missing data condition and

where home educational resources turns out to be a significant predictor of math achievement ($B = 0.87$, $p = .048$). This observation of an exception underscores the importance of relying on more than one missing data handling method when percentage of missing data is large (exceeds 5%) as also suggested by Raymond and Roberts (1997).

Although the R^2 values presented in Tables 2 and 3 suggest that regression imputation and multiple imputation methods provide effect size estimates that closely match their full data counterparts, it should be noted that the resulting gains in efficiency are very small compared to listwise deletion (< 1%). In other words, for the large sample ($n = 5{,}455$) used in this example, listwise deletion is almost as good a choice as the best missing data imputation method. The next step is to see if this result also holds when the sample size is relatively much smaller.

Empirical Sample 2

For the small sample illustration U.S. Census Bureau (2000) data were used. This dataset was used to test for mean difference in percentage of individuals, twenty-five years or older, with college degrees at county level between the states of Virginia and Wisconsin. The sample size was 207 (Virginia, $n = 135$; Wisconsin, $n = 72$). The independent samples t test results based on various missing data handling methods are presented in Tables 4 and 5 under low, $m = 5\%$ ($n = 197$) and high, $m = 10\%$ ($n = 186$) missing data conditions respectively.

These results show that, in terms of effect size, best results are obtained with listwise deletion ($d = .26$) and ML imputation ($d = .26$) when the proportion of missing data is small, and with mean imputation ($d = .25$) when the proportion of missing data is large. Power statistics suggest a small increases in power, from .915 to .926 (gain = 1.2%) when proportion of missing data is small and from .894 to .926 (gain = 3.8%) when proportion of missing data is large. In terms of the effect on test statistic, results were not consistent for all missing data handling methods. For instance, with 5% missing data the null hypothesis of no significant mean difference in percentage of individuals, twenty-five years or older, with college degrees at county level between the states of Virginia and Wisconsin was rejected under listwise deletion ($t = 2.08$, $p = .039$), mean imputation ($t = 2.19$, $p = .030$), ML imputation ($t = 2.09$, $p = .038$), and multiple imputation ($t = 1.87$, $p = .038$), but not under regression imputation ($t = 1.84$, $p = .067$). With 10% missing data, this same null hypothesis was rejected under mean imputation ($t = 2.02$, $p = .044$) and regression imputation ($t = 2.18$, $p = .031$) but not under listwise deletion ($t = 1.82$, $p = .071$), ML imputation

Table 4. Independent samples *t* test results for education attainment with 5% missing data under various missing data handling methods using the 2000 census data

	Summary statistics					
	Virginia			*Wisconsin*		
	n	*M*	*SD*	*n*	*M*	*SD*
Full data	135	19.70	11.47	72	17.22	6.16
Listwise deletion	128	19.87	11.38	69	17.26	6.28
Mean imputation	135	19.87	11.08	72	17.26	6.14
Regression imputation	135	19.67	11.17	72	17.42	6.36
EM imputation	135	19.83	11.08	72	17.33	6.15
Multiple imputation	135	19.76	11.35	72	17.17	6.44

	***t* test statistics**						
	t	*df*	*p*	Δ*M*	*SE*(Δ*M*)	*d*	*Power*
Full data	2.02*	204.98	0.045	2.47	1.23	0.25	0.926
Listwise deletion	2.08*	194.88	0.039	2.62	1.26	0.26	0.915
Mean imputation	2.19*	204.66	0.030	2.62	1.20	0.27	0.926
Regression imputation	1.84	204.08	0.067	2.25	1.22	0.23	0.926
EM imputation	2.09*	204.64	0.038	2.50	1.20	0.26	0.926
Multiple imputation	1.87*	204.08	0.038	2.59	1.24	0.21	0.926

Note. *n* = 207. *df* = degrees of freedom. The *t* and *df* values are reported after adjustment for unequal sample sizes and unequal group variances. Δ*M* = mean difference. *d* = Cohen's *d*; * *p* < .05; ** *p* < .01; *** *p* < .001

Table 5. Independent samples *t* test results for education attainment with 10% missing data under various missing data handling methods using the 2000 census data

	Summary statistics					
	Virginia			*Wisconsin*		
	n	*M*	*SD*	*n*	*M*	*SD*
Full data	135	19.70	11.47	72	17.22	6.16
Listwise deletion	128	19.87	11.38	63	17.28	6.23
Mean imputation	135	19.87	11.08	72	17.28	5.82
Regression imputation	135	19.67	11.17	72	16.83	6.10
EM imputation	135	19.83	11.08	72	17.48	5.84
Multiple imputation	135	19.76	11.35	72	17.37	6.64

	***t* test statistics**						
	t	*df*	*p*	Δ*M*	*SE*(Δ*M*)	*d*	*Power*
Full data	2.02*	204.98	0.045	2.47	1.23	0.25	0.926
Listwise deletion	1.82	183.57	0.071	2.37	1.31	0.23	0.894
Mean imputation	2.02*	204.98	0.044	2.37	1.17	0.25	0.926
Regression imputation	2.18*	204.96	0.031	2.63	1.21	0.27	0.926
EM imputation	1.79	204.99	0.075	2.11	1.18	0.22	0.926
Multiple imputation	1.87	202.46	0.071	2.37	1.27	0.21	0.926

Note. *n* = 207. *df* = degrees of freedom. The *t* and *df* values are reported after adjustment for unequal sample sizes and unequal group variances. Δ*M* = mean difference. *d* = Cohen's *d*; * *p* < .05; ** *p* < .01; *** *p* < .001

($t = 1.79$, $p = .075$), and multiple imputation ($t = 1.87$, $p = .071$). These contradictory results stand in sharp contrast to results of tests of hypothesis obtained earlier in example 1 and underscore the risks inherent in using any missing data handling method when a large proportion of data is missing in a small sample.

Discussion

The primary objective of this study was to formulate general guidelines that can assist educational researchers in the selection of appropriate missing data handling methods under various combinations of sample size, proportion of missing data, and analytical method. By keeping all of these factors constant, any observed differences in performance of missing data handling methods can more or less be attributed to the relative efficiency of those methods. The statistical analyses conducted in this study can be thought of as a response to recommendations made in earlier studies such as Roth (1994) and Young et al. (2011) who identified a need for guidelines that can help researchers choose missing data handling methods under a variety of scenarios.

Although previous research exists that has looked at the effect of factors such as sample size, proportion of missing data, and method of analysis on the effectiveness of missing data handling methods, there are no clear cut guidelines which can inform a researcher as to which missing data handling method is best under which circumstances. Prior studies used different samples with varying proportions of missing data under different analytical methods which makes it very difficult to isolate the effect of any single factor. The present study is an attempt to rectify this state of affairs. It is hoped that insights provided by the findings of this study will further publicize the issues involved and encourage further research in this direction.

In some respects the present study has been able to confirm and support earlier findings. For example, our statistical results imply that listwise deletion is one of the simplest, easily justified, and least computation-intensive methods under large sample and low missing data conditions when the objective is to obtain consistent and unbiased estimates of population parameters (Haitovsky, 1968; Wayman, 2003; Young et al., 2011). On the other hand, the use of this method comes at the price of sacrificing additional statistical power that can be gained by imputing missing data. One can make a case that if sample size is large enough such that achievement of adequate power is not a concern, then listwise deletion provides one of the least risky (since it avoids adding another layer of measurement error to the data) and most quickly deployable missing data handling methods. Even in

cases where listwise deletion is not the best missing data handling method, for instance in terms of efficiency, it still remains an attractive choice because the efficiency gains offered by competing methods are often trivial making it difficult to justify the increased computational complexity in statistical analyses due to their employment.

We further confirmed the general finding of past studies that if missing data imputation is unavoidable, then the two best methods for such imputation are maximum likelihood imputation (e.g. Expectation-maximization imputation) and multiple imputation (Graham et al., 1996; Wayman, 2003; Peugh and Enders, 2004; Peng et al., 2006; Young et al., 2011; Knol et al., 2010). This can be clearly seen from the figures presented in Table 1 which show that ML and multiple imputation methods performed best in 22 out of 24 (91.6%) scenarios depicted therein. In order to get a more complete ranking of the five missing data handling methods used in this study, we used a simple scoring method where the least-performing to best-performing methods received a score from 1 to 5 for each of the 120 possibilities based on sample size (small, medium, large), proportion of missing data (low, high), the five missing data handling methods, and the four methods of analysis. The sum of scores across missing data handling methods revealed the following ranking and total scores: multiple imputation, 104; expectation maximization, 83; listwise deletion, 65; regression imputation, 63; and mean imputation, 45.

Although listwise deletion is in the third place in this ranking we reiterate our earlier contention that it is often preferable over other methods when the gain in estimation accuracy offered by those methods is trivial. This ranking of missing data handling methods also makes intuitive sense as it ranks these methods in the order of their mathematical sophistication, ranging from the most sophisticated, multiple imputation which offers most realistic modeling of random variation, to the least sophisticated, mean imputation method that offers no accommodation for random variability.

The important thing to note here is that the positive effect of gain in accuracy of parameter estimates due to missing data imputation does not always dominate the negative effect of measurement error introduced by such imputation. For instance, our results showed that in many instances listwise deletion, that is the no imputation method, worked better than some imputation methods but not others even after controlling for method of analysis, sample size, and proportion of missing data. For example, in our simulation two-way ANOVA for a medium sample with high proportion of missing data, listwise deletion performed better than mean imputation but worse than multiple imputation. For mean imputation in this scenario the positive effect of missing data imputation was dominated by the

negative effect of larger measurement error due to that imputation. On the other hand, the reverse was true for multiple imputation where the positive effect of missing data imputation dominated the negative effect of larger measurement error due to such imputation. The message here is that missing data imputation is not always an improvement over non-imputation and that some missing data imputation methods can actually cause more harm than benefit.

An important implication of our statistical results is that missing data imputation can be beneficial in raising the statistical power of tests of hypothesis. In our simulated data relative power gain ranged between 0% and 28.8% while absolute power gain ranged between 0 and .12, depending on sample size, proportion of missing data, and method of analysis used. The gains in statistical power were pronounced for small samples, $n \leq 50$, in general (min gain = .003 or 0.4%; max gain = .11 or 28.8%; mean gain = .03 or 10.4%) and for small samples with high proportions of missing data ($m > 5\%$) in particular (min gain = .003 or 2.87%; max gain = .11 or 28.8%; mean gain = .04 or 14.9%). For sample sizes exceeding 200, statistical power was not an issue for any of the four methods of analysis adopted in this study (min power = .98; max gain = .01 or 1.2%). Similarly the gains in power were modest when proportion of missing data was 5% or less (max gain = .03 or 6.7%). The bottom line here is that statistical power by itself can be an important consideration for choosing missing data imputation even in cases where the non-missing pre-imputation data represents the target population well and listwise deletion is a viable option. This is especially true for small samples with large proportions of missing data.

The importance of statistical power issues highlighted in the preceding paragraphs should not be taken to mean that population representation is a minor consideration. Even when sample size is large and statistical power is not an issue, the occurrence of missing data can transform the sample in such a way that it is no longer representative of its target population. In such cases it is important to impute missing data or alternately, if possible, to use adjusted sampling weights in order to make the sample representative again. One may argue that the use of sampling weights is preferable over missing data imputation because the former method does not introduce additional measurement error.

Recommendations Based on Sample Size

Regarding choice of missing data handling method our general recommendation is that if (1) sample size is large enough for adequate power, and (2) sample is representative of the target population, then use listwise deletion. In cases where either of these conditions is not met the best methods are multiple imputation and maximum likelihood imputation. It is important to note here that these recommendations are for missing data that are either missing at random (MAR) or missing completely at random (MCAR), and not for data that are not missing at random (NMAR).

When sample size is large, $n \geq 1{,}000$, lack of statistical power is generally not an issue as clearly demonstrated by our simulated results and empirical data examples. The decision to impute missing data thus depends on whether or not the non-missing data are still representative of the target population. For small samples, in terms of gain in accuracy of estimation, the best available methods of missing data imputation are maximum likelihood imputation and multiple imputation. Although strictly speaking multiple imputation on average performs better than ML imputation in small samples we recommend using more than one imputation method in general when the sample size is small and in particular when sample size is small and proportion of missing data is high in order to lower the risk of getting into the unfortunate situation where the negative effect of an increase in measurement error due to imputation exceeds the positive effect of a gain in estimation accuracy due to that imputation.

Our recommendations for choice of missing data handling method are summarized in Figure 2. If the missing data are MCAR and the resulting sample after listwise deletion provides adequate power for tests of hypotheses, then listwise deletion should be used. If the missing data are MAR, then listwise deletion should only be used if the resulting sample after listwise deletion is still representative of the population and there is adequate power for tests of hypotheses. Finally, if missing data is NMAR, then the missing data mechanism must be modeled as part of the estimation process. Because the term NMAR is an umbrella term for all sorts of non-random missing data mechanisms, the exact modeling process depends on the type of non-randomness present in the missing data. For example, if the missingness is due to selection bias, Heckman correction can be used (Heckman, 1979). We recommend multiple imputation and maximum likelihood imputation as the methods of choice.

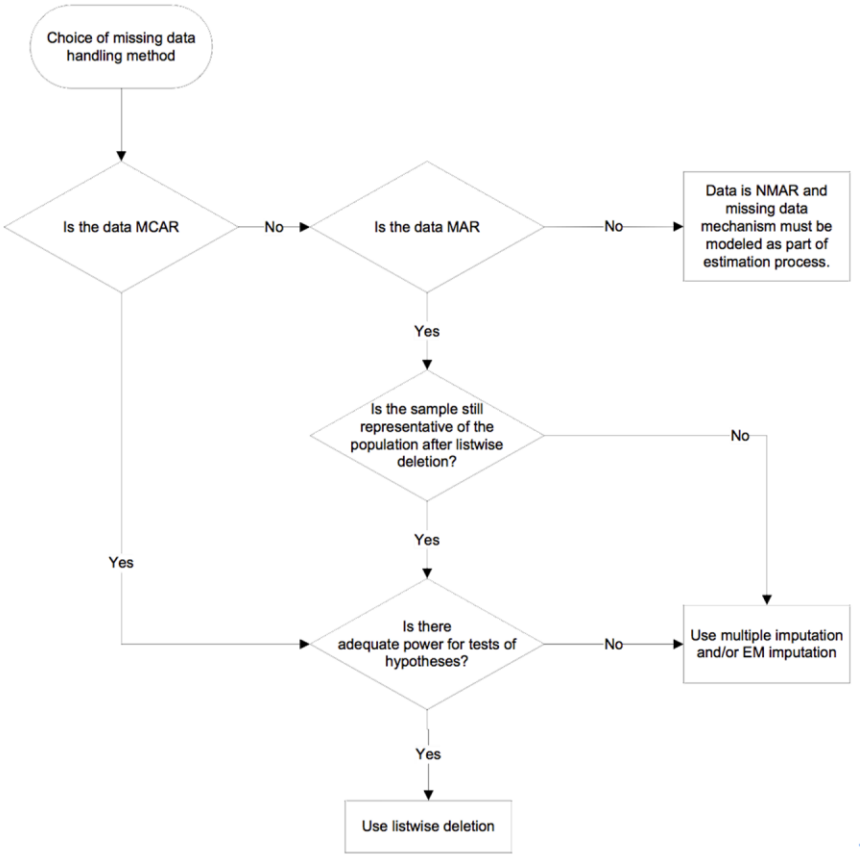

Figure 2. The decision process governing choice of missing data handling method

Scope for Future Research

There are several directions for future research. First, more work needs to be done on the effect of missing data handling methods on method of analysis. All four methods of analysis adopted for statistical analyses presented in this study, one sample *t* test, independent samples *t* test, two-way ANOVA, and multiple regression, are special cases of the general linear model. It would be interesting to see whether the guidelines developed here are also applicable to nonlinear models, for example models of count data such as logistic regression. There is also further scope for testing these guidelines in context of longitudinal, repeated measures, and multi-level models.

The second potential line of research is to focus on application. Future studies can take an applied approach and use real-life datasets from various subfields of

education in order to evaluate the effectiveness of guidelines presented in this study. The importance of simulation work notwithstanding, it is the presence or lack of empirical evidence which is most important in determining whether or not such guidelines may see widespread acceptance in educational research.

References

Acock, A. (2005). Working with missing values. *Journal of Marriage and Family, 67*(4), 1012-1028. doi:10.1111/j.1741-3737.2005.00191.x

Afifi, A., & Elashoff, R. (1966). Missing observations in multivariate statistics: I. Review of the Literature. *Journal of the American Statistical Association, 61*(315), 595-604. doi:10.2307/2282773

Allison, P. (2001). *Missing data*. (Sage University Paper Series on Quantitative Applications in the Social Sciences, series no. 07-136). Thousand Oaks, CA: Sage. doi:10.4135/9781412985079

Alosh, M. (2009). The impact of missing data in a generalized integer-valued autoregression model for count data. *Journal of Biopharmaceutical Statistics, 19*, 1039–1054. doi:10.1080/10543400903242787

Andridge, R., & Little, R. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review, 78*(1), 40-64. doi: 10.1111/j.1751-5823.2010.00103.x

Brick, J., & Kalton, J. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research, 5*, 215-238. doi:10.1177/096228029600500302

Cohen, J. (1992). Quantitative methods in Psychology: A power primer. *Psychological Bulletin, 112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Graham, J., Hofer, S., & MacKinnon, D. (1996). Maximizing the usefulness of data obtained with planned missing value patterns: An application of maximum likelihood procedures. *Multivariate Behavioral Research, 31*(2), 197-218. doi:10.1207/s15327906mbr3102_3

Groves, R., Dillman, D., Eltinge, J., & Little, R. (2002). *Survey Nonresponse*. New York, NY: John Wiley & Sons, Inc.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2004). *Survey methodology*. Hoboken, NJ: John Wiley & Sons, Inc.

Haitovsky, Y. (1968). Missing data in regression analysis. *Journal of the Royal Statistical Society, Series B (Methodological), 30*(1), 67-82.

Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica, 47*(1), 153-161. doi:10.2307/1912352

Jones, M. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91*(433), 222-230. doi:10.2307/2291399

Knol, M., Janssen, K., Donders, A., Egberts, A., Heerdink, E., Grobbee, D., Moons, K., & Geerlings, M. (2010). Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *Journal of Clinical Epidemiology, 63*, 728-736. doi:10.1016/j.jclinepi.2009.08.028

Madow, W., Nisselson, H., & Olkin, I. (Eds.). (1983). *Incomplete data in sample surveys, Volume 1: Report and case studies.* New York, NY: Academic Press.

Madow, W., & Olkin, I. (Eds.). (1983). *Incomplete data in sample surveys, Volume 3: Proceedings of the symposium.* New York, NY: Academic Press.

Madow, W., Olkin, I., & Rubin, D. (Eds.). (1983). *Incomplete data in sample surveys, Volume 2: Theory and bibliographies.* New York, NY: Academic Press.

National Center for Education Statistics. (2003). *Program for International Student Assessment* [Data file]. Retrieved from http://nces.ed.gov/surveys/pisa/datafles.asp

Peng, C., Harwell, M., Liou, S., & Ehman, L. (2006). Advances in missing data methods and implications for educational research. In S. S. Sawilowsky (Ed.), *Real Data Analysis* (pp. 31-78). Charlotte, NC: New Information Age Publishing.

Peugh, J., & Enders, C. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525-556. doi:10.3102/00346543074004525

Raymond, M. & Roberts, D. (1987). A comparison of methods for treating data in selection research. *Educational and Psychological Measurement, 47*(1), 13-26. doi:10.1177/0013164487471002

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. *Personnel Psychology, 47*(3), 537-560. doi:10.1111/j.1744-6570.1994.tb01736.x

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons, Inc. doi:10.1002/9780470316696

Schafer, J., & Graham, J. (2002). Missing data: Our view on the state of the art. *Psychological Methods, 7*(2) 147-177. doi:10.1037//1082-989X.7.2.147

U.S. Census Bureau. (2000). *Census of population and housing* [Data file]. Retrieved from http://www.ers.usda.gov/data/education

Wayman, J. (2003, April). *Multiple imputation for missing data: What is it and how can I use it?* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Young, W., Weckman, G., & Holland, W. (2011). A survey of methodologies for the treatment of missing values within datasets: limitations and benefits. *Theoretical Issues in Ergonomics Science*, *12*(1), 15-43. doi:10.1080/14639220903470205