HANDLING MISSING DATA IN EDUCATIONAL RESEARCH USING SPSS

by

Jehanzeb Cheema

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Education

Committee:

Dimitrov _____ Chair

_____

_____

_____ Program Director

_____ Dean, College of Education and Human Development

Date: April 3, 2012 _____ Spring Semester 2012
George Mason University
Fairfax, VA

Handling Missing Data in Educational Research Using SPSS

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Jehanzeb Cheema
Doctor of Philosophy
University of Wisconsin-Milwaukee, 2006
Master of Arts
University of Wisconsin-Milwaukee, 2003
Bachelor of Science
Slippery Rock University of Pennsylvania, 2002
Bachelor of Arts
University of the Punjab, 1999

Director: Dimiter Dimitrov, Professor
Graduate School of Education

Spring Semester 2012
George Mason University
Fairfax, VA

Dedication

To Sahar.

Acknowledgements

I am thankful to Allah Sub'hanahu wa Ta'ala who gave me the ability to write this dissertation. I will also like to acknowledge my parents and parents-in-law for their support, Prof. Dimiter Dimitrov for supervising my research, Prof. Anastasia Kitsantas for her guidance, Dr. Herbert Ware for his precious comments and all of my former professors and instructors who contributed to my learning.

Table of Contents

List of Tables

List of Figures

Abstract

HANDLING MISSING DATA IN EDUCATIONAL RESEARCH USING SPSS

Jehanzeb Cheema, Ph.D.

George Mason University, 2012

Dissertation Director: Dr. Dimiter Dimitrov

This study looked at 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, in order to evaluate the effect of missing data treatment on accuracy of estimation. In order to accomplish this 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.

1. 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). 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 analytical procedures directly depends on the quality of underlying data which in turn can be affected by the nature of missing data (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 is handling missing data can vary considerably depending on that researcher's circumstances.

This study is not the first one to investigate or compare the performance of missing data handling methods. Such methods have been investigated in the past both in general (Afifi & Elashoff, 1966; Graham, Hofer, MacKinnon, 1996; Haitovsky, 1968; Peng, Harwell, Liou, & Ehman, 2009; Peugh & Enders, 2004; Wayman, 2003; Yesilova, Kaya, Almali, 2011; 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). However, 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 other datasets. The current study not only controls for all of these factors simultaneously but compared to past studies 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.

The main objective of this study is to provide educational researchers with general guidelines about which missing data imputation 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. A second objective of this study is to provide guidelines to educational researchers about the effect of missing data imputation on statistical power of tests of hypotheses.

In term of research questions, this study specifically asks (1) which missing data handling method works best when method of analysis, proportion of missing data, and sample size are known? and (2) What is the effect of missing data imputation on statistical power of tests of hypotheses?

List of Important Terms and Definitions

This section contains definitions of basic terms used throughout the remaining chapters of this study. Some of these terms are explained in more detail in other chapters.

Missing completely at random (MCAR). MCAR is the missing data mechanism when missing data for a variable is not related to other variables in the dataset or to the values of that variable itself. For example, if *Y* (say, self-efficacy) is unrelated to any of the predictors in the dataset (say, gender and race) and when missing data on *Y* is not related to the value of *Y* itself (i.e. we do not have a situation where more people with high or low self-efficacy values are non-respondents), then missing data on *Y* is said to be MCAR (Allison, 2001).

Missing at random (MAR). MAR is the missing data mechanism when missing data for a variable is related to other variables in the dataset but not to the values of that variable itself. For example, under the MAR assumption, the missing data for *Y* (say, self-efficacy) may depend on another variable *X* (say, race) but is not related to the value of *Y* when *X* is controlled for. A counter example is that of salary where salary may be

related to race but even after controlling for race, missing data on salary may still be related to the value of salary itself, for example when individuals with higher salaries are reluctant to report their salaries (Allison, 2001).

Not missing at random (NMAR). This is the missing data mechanism whenever data is neither MCAR nor MAR. In other words, NMAR comprises of all situations that are not covered under the MCAR and MAR mechanisms. An example of NMAR variable is salary. Individuals with high salaries tend to not report their salaries as compared to those who have lower salaries. Thus, the probability of a missing value for salary is a function of the value of salary itself. The missing data in this case is thus NMAR (Allison, 2001).

Root mean squared error (RMSE). RMSE is a measure of accuracy that is calculated as the square root of average squared difference between observed and true values of a parameter. For example, if the observed values of a parameter in three difference samples are 1, 5, and 5, and the true value is 3, then RMSE is,

$$RMSE = \sqrt{\frac{(1-3)^2 + (5-3)^2 + (5-3)^2}{3}} = 2$$

Thus, *RMSE* is in essence the average distance of observed error from the true value and can be roughly interpreted as the standard deviation of $X_{Observed}$ from their true value rather than the sample mean. This measure thus takes into consideration the absolute size of error (Weisstein, 1999). 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. The normalized root mean squared error for m observations of $X_{Observed}$ that have a true value of X_{True} can be expressed as follows (Taguchi et al., 2009).

$$Normalized\ RMSE = \frac{\sqrt{\frac{\sum_{i=1}^{m}\left(X_{Observed,i} - X_{True}\right)^2}{m}}}{Max(X) - Min(X)}$$

In our example, the RMSE value of 2 can be normalized by dividing with the range of observed values,

$$Normalized\ RMSE = \frac{2}{5-1} = 0.5$$

Maximum likelihood (ML). Maximum likelihood is a mathematical procedure that can be used to find one or more parameters of a statistical model that, for the observed data, maximize the observed likelihood distribution. For example, if a process follows the binomial distribution with say $n = 10$ trials and $x = 9$ observed instances of success, then we can simply input these observed data into the probability distribution function (the formula for binomial distribution in this example), and maximize that function with respect to the values of the unknown parameters. Thus, with $n = 10$ trials and $x = 9$ successes, the binomial distribution function becomes,

$$P(x=9) = \binom{10}{9} p^9 (1-p)^{10-9} = 10p^9(1-p) = 10p^9 - 10p^{10}$$

In the preceding expression, p is the probability of success. It can be easily verified that for our observed data with $n = 10$ trials and $x = 9$ successes, the binomial

distribution function is maximized when $p = 0.9$ with $P(x = 9) = .387$. At all other values of p, the likelihood function is less than maximum (the reader can easily verify this by substituting alternative values of p in the binomial formula with $n = 10$ and $x = 9$). In this example, the value $p = 0.9$ is called the maximum likelihood estimate of parameter p (Agresti, 2007). In other words, this is the value of parameter p at which the probability of the observed data is maximum (hence the name, maximum likelihood).

Expected value of a statistic. The expected value of a statistic is the weighted average of all possible values that this statistic can take. For example, the expected value of arithmetic mean, $E(X)$, is the average value of arithmetic mean from all possible samples that can be derived from a given population. Another way to think about the expected value of a statistic is that it is the average value of the statistic that we expect to obtain when an experiment is repeated a very large number of times (Howell, 2007).

Standard error of a statistic. Standard error of a statistic is the standard deviation of all possible values of a statistic of interest. For example, assume that for a given population we are interested in the statistic, sample mean. Sample means can be calculated for all possible samples that can be drawn from that population. Standard error of the sample mean is then nothing but the standard deviation of all the sample means calculated in this way (Howell, 2007).

Monet Carlo simulation. Monet Carlo simulation is a method that is often used as an alternative to analytical proof in cases where such proof is either computationally too complex to build or is impossible to build (for instance, due to the presence of random probabilistic elements). The Monet Carlo simulation method involves drawing

samples repeatedly a large number of times (often hundreds or thousands of times) and calculating statistics of interest from all those samples. Results from all such samples are then averaged to provide long-run or expected values of those statistics (Gujarati, 2003).

Expectation-maximization (EM) algorithm. The EM imputation is a maximum likelihood-based iterative method that involves two steps. In the first step initial values (often marginal means) are assigned to missing data. In the second step, expectations formed with those initial values are maximized. This expectation-maximization cycle is then repeated again and again until the imputed values converge based on a pre-determined convergence criteria. The EM imputation method produces unbiased standard errors of parameter estimates (Salkind & Ramussen, 2007).

2. Literature Review

Missing data are encountered regularly by researchers in educational research. Most large scale, especially nationally representative, educational datasets in the U.S. contain thousands of individual cases. Unfortunately, such data is seldom complete. Presence of missing data on one or more variables of interest for a proportion of the sample has become a rule rather than an exception in large scale survey research (McKnight, McKnight, Sidani, & Figueredo, 2007; Peng et al., 2006). A study that contains many variables with a relatively small number of missing values can cause significant attrition in the total effective sample size. For example, a dataset containing 500 observations and 10 variables with 10% of the data independently missing on each variable can reduce the effective sample size with listwise deletion to just $.9^{10} \times 500 = 175$. For many methods of analysis such attrition in sample size can force the researcher to choose alternative methods due to the fall in power as a result of reduction in *n*. There are various reasons why data may be missing in surveys. Sometimes it is because the respondents intentionally ignore certain questions. For example, a respondent may not feel comfortable answering questions about her salary or her criminal record. In some cases a respondent may genuinely forget to answer a specific question or an interviewer may forget to ask a question. Other reasons for missing data include the inapplicability of a certain question to the respondent or the inability of the respondent to

answer a question, for example, due to the respondent's death in a longitudinal study (Allison, 2001; Groves et al., 2004). The current literature offers many missing data imputation methods ranging from very simple, such as mean (or median) imputation where missing data on a variable are substituted simply by the mean (or median) of the non-missing data, to fairly complex procedures such as expectation-maximization which assigns random initial values to missing data and then proceeds to maximize the expectations formed with those initial values in an iterative sequence (Dempster, Laird, & Rubin, 1977). Past studies that dealt with imputation of missing data go as far back as the 1930's. For example, Wilks (1932) proposed a maximum likelihood method for imputation of missing data in bivariate normal distributions. With the availability of industrial scale computing, a surge in interest related to missing data imputation occurred during the 1950's and 1960's (Afifi & Elashoff, 1966; Buck, 1960; Edgett, 1956). During these decades several advanced methods for handling missing data, such as linear regression, were introduced. However, lack of statistical packages that could deliver advanced imputation methods and the scarcity of computing resources at the disposal of individual researchers meant that there was little progress in this field during that time. Early 1980's and 1990's saw the start of a renewed interest in missing data imputation thanks to now accessible statistical packages that could easily implement such methods and widespread access to computing resources. These developments allowed a large number of researchers unprecedented access to large-scale datasets and as a result there was a surge in the number of new missing data imputation methods (Brick & Kalton, 1996; Hong & Wu, 2011; Zhou, Wang, & Dogherty, 2003).

Sources and Consequences of Missing Data

Brick and Kalton (1996), and Groves et al. (2004) identify three principal sources of missing data in survey research: non-coverage, total non-response, and item non-response. Missing data due to non-coverage occurs when some population units are left outside the sampling frame and thus have no chance of being selected in the sample. Missing data due to total non-response occur when a respondent refuses to respond to any item on the survey i.e. the entire row in the dataset representing that respondent has missing data. Item non-response occurs when a respondent responds to only a fraction of items on the survey. Of these three sources of missing data, non-coverage and total non-response can be fixed by using appropriate sampling weights that are designed to make the sample accurately represent the target population. Missing values due to item non-response can however not be fixed by using weights. The choice is thus between listwise deletion of cases with item non-response, which results in loss of some of that valuable information that those cases did provide, and missing value imputation, which introduces an additional layer of error in parameter estimation because such imputed data, however precisely imputed, is unlikely to exactly match the missing information. Brick and Kalton (1996) identify another source of missing data, partial non-response, where respondents provide responses to only a very small number of items. However, this kind of non-response falls in-between total non-response and item non-response and can be corrected by either using weights or missing data imputation depending on how the researcher wants to treat such non-response. Thus, whenever missing data due to item non-response is encountered by a researcher, a trade-off needs to be made by comparing

the net benefit of more precision at the expense of losing some information with that of using imputed data at the cost of a potentially larger measurement error.

Although one of the concerns with missing data is the attrition in sample size, even in cases where such attrition is not large, the concern remains about whether the remaining sample is still representative of the target population. Since the nature and properties of missing data can be very different from observed data, it is important to analyze various methods of treating missing data in order to determine which methods work best under a given set of conditions. The simplest situation is when the missing data can be completely ignored. This is a legitimate strategy when the reduced sample size due to missing data still accurately represents the target population. Rubin (1976) terms such missing data as missing completely at random (MCAR). An example of MCAR data is when missing data on a variable *Y* (say, self-efficacy) is unrelated to any of the predictors in the dataset (say, gender and race) and when missing data on *Y* is not related to the value of *Y* itself (i.e. we do not have a situation where more people with high or low self-efficacy values are non-respondents). When data is MCAR it allows the researcher to perform data analysis procedures on the reduced sample as if it were the full sample without any lack of generalizability. However, if the sample obtained after discarding missing data no longer represents the population of interest, then any findings based on that sample will not be generalizable to the population thus severely restricting the usefulness of such findings. In such a situation, discarding missing data from the original sample will obviously not suffice and in order to have parameter estimates that are consistent and unbiased, the researcher must resort to missing data imputation.

Finally an important point to remember is that even in situations where case deletion methods, such as listwise deletion, produce samples that represent their corresponding population well, power of the analysis does get reduced due to the positive relationship between sample size and power. Thus, even in cases where listwise deletion may produce consistent and unbiased estimates pf parameters, it may be desirable to use a more sophisticated missing data handling method in order to conserve power.

Missing Data Mechanisms: MCAR, MAR, and NMAR

The appropriateness of a missing data handling method is contextual and depends on the missing data mechanism. One such mechanism, MCAR, has already been discussed. When data is not MCAR, it can either be missing at random (MAR) or not missing at random (NMAR). The data is MAR when the probability of missing data on a variable is unrelated to the value of that variable itself but may be related to the values of other variables in the dataset. For example, under the MAR assumption, the missing data for *Y* (say, self-efficacy) may depend on another variable *X* (say, race) but is not related to the value of *Y* when *X* is controlled for. A counter example is that of salary where salary may be related to race but even after controlling for race, missing data on salary may still be related to the value of salary itself, for example when individuals with higher salaries are reluctant to report their salaries.

The data is NMAR when the probability of missing data on a variable is a function of the value of that variable itself (Rubin, 1976; Allison, 2001). An example of NMAR variable is salary. Individuals with high salaries tend to not report their salaries as compared to those who have lower salaries. Thus, the probability of a missing value for

salary is a function of the value of salary itself. The missing data in this case is thus NMAR. The MAR assumption is a less restrictive form of MCAR assumption, the latter assumption essentially being that the probability of missing data on a variable should not be related to the value of that variable itself or to the values of other variables in the dataset. When the data is either MCAR or MAR, there is no need to model the missing data mechanism as part of the estimation process. In other words, once the missing data handling method has been applied to MCAR and MAR data, any method of analysis can be used with the resulting dataset as if it were complete. When the data is NMAR, the missing data mechanism needs to be specifically modeled as a part of the estimation process due to the fact that for NMAR data the parameter estimates of the method of analysis are not independent of the process through which data is missing. Imputation of NMAR data requires extensive *a priori* knowledge of the missing data process as such process cannot be determined from the observed data itself. For these reasons, missing data handling methods for NMAR data must be tailored in context of the missing data process and cannot be used to construct general guidelines that are applicable under relatively stronger assumptions of MCAR and MAR (Allison, 2001).

At this point it is natural to ask what happens if we treat these mechanisms incorrectly. When NMAR data is incorrectly treated as MCAR or MAR, it means that we are specifically not modeling the missing data process and that parameter estimates of our method of analysis will not be correct. When MCAR data is incorrectly treated as MAR or NMAR, it means that we are unnecessarily introducing more complexity in our handling of missing data by giving up the simple method of listwise deletion, that can

generate unbiased and consistent estimates of our model parameters, in favor of an imputation method and/or in favor of modeling the missing data process, either of which, if implemented improperly, can result in parameter estimates that are inferior to those obtained from listwise deletion. The case of treating MAR data as NMAR is the same as that of treating MCAR data as NMAR. Finally, when MAR data is incorrectly treated as MCAR, it means that we are over-simplifying the handling of missing data and even though we can still generate parameter estimates, those estimates will not be generalizable to the population (Allison, 2001).

Ignorable Versus Nonignorable Missing Data

In missing data research, the MAR mechanism is closely related to missing data mechanism referred to as ignorable (Allison, 2001; Rubin, 1976). Missing data is ignorable when such data is (1) MAR and (2) the parameters that are responsible for occurrence of missing data are unrelated to the parameters being estimated. Since in real-world situations the latter assumption is almost always satisfied, the MAR mechanism can be thought of as ignorable (Allison, 2001). In contrast to the ignorable mechanism, missing data are nonignorable when it is expected that cases with missing data on a variable have significantly different (higher or lower) values on that variable than the cases with complete information, after controlling for the effect of all remaining variables in the data. An example of nonignorable data is when pretest, posttest and follow-up scores are collected in a longitudinal study. Such data would be considered MAR if the dropouts from the study were related to only those variables that were collected at occasions prior to such dropout (Schafer & Graham, 2002). The missing data mechanism

would be considered nonignorable if data collected after participants dropped out of the study were in some way related to those dropouts. For example, in a survey of eating habits, if all obese respondents drop out of the study before the posttest phase, then the nonresponse in posttest phase is unrelated to variables obtained at the pretest phase. Now, if the eating habits of obese and non-obese individuals actually differ from each other, then the missing data mechanism is nonignorable.

Although the best way to deal with missing data is to look at each dataset individually and determine the requirements for imputing its missing data based on the specific features of that dataset, educational researchers often do not possess the expertise required to identify and implement the best imputation method applicable to their specific requirements (McKnight et al., 2007) and often end up using simpler but easier-to-implement ready-made imputation methods provided by popular software packages, sometimes not even realizing that the use of an incorrect method can introduce serious bias in their estimation results (Allison, 2001; Wayman, 2003). The reasons for not possessing sufficient expertise to understand and apply advanced statistical methods vary among educational researchers. Murtonen and Lethtinen (2003), for example, identify factors such as superficial teaching received, difficulty in linking theory with practice, difficulty of content involved, inability to visualize an integrative picture of research, and negative attitude towards statistics as major hurdles in the learning of statistical concepts in education and sociology. These issues are especially relevant to a majority of graduate students in education who specialize in areas other than quantitative research methods, and have taken only a few basic (usually compulsory) courses in quantitative methods.

This leaves such researchers without the expertise to evaluate the appropriateness of a particular missing data handling method in context of their own research.

As noted earlier, several past studies have compared missing data handling techniques in order to identify methods that provide the highest degree of accuracy in parameter estimation. Some studies have also looked at the effect of sample size and the proportion of missing data on the effectiveness of missing data imputation method. It is appropriate to introduce the various missing data handling methods before summarizing key findings of those studies.

Missing by Design

Up to this point we have only considered instances where data is missing unintentionally. In other words, we have not discussed a situation where data is allowed to be missing as a feature of the study design. In some cases the design of the study exposes one group of participants to one set of variables and a second group of participants to a different set (McKnight et al., 2007). For example, different groups of students taking the same standardized test (such as SAT) may be exposed to different sets of items. In such cases, for a particular item, responses are not available for all students. However, the performance of students exposed to different sets of items can still be compared with each other by either mapping a one-to-one equivalence between items (for instance two different items that are rated as equal in terms of difficulty and content can be considered alternative versions of the same question), or by using advanced methods such as those proposed by item response theory. Situations where data is missing by design frequently occurs in longitudinal studies, in studies that involve procedures that

impose a large burden on the respondent (for example in terms of time commitment or due to the involvement of painful procedures), and in single-case designs where efficient study designs eliminate the need for complete data (Kennedy, 2005). For example, depending on the context, in a study involving two subjects, an A-B-C design for subject 1 and an A-B-A design for subject 2 may provide more information than an A-B-A design for both participants.

Missing Data Patterns

It is not always possible to determine the mechanism underlying missing data. One exception is MCAR data for which Little's MCAR test (Little, 1988) exists which is based on the missing data patterns identifiable from observable data. Little's MCAR test is implemented as a chi-squared test in SPSS with the null hypothesis that missing data is MCAR. The term missing data pattern here refers to sorting the dataset into groups based on whether a case has missing or non-missing value on a certain variable. For example, in a bivariate dataset containing missing data for variables *X* and *Y*, four possible patterns are possible for a given case: Both *X* and *Y* values are missing, only *X* value is missing, only *Y* value is missing, neither *X* nor *Y* is missing. Take another example where *Y* is achievement score and *X* is gender. One may analyze missing data with respect to gender in order to see whether the missing data does or does not vary between males and females. If 20% of achievement scores for males are missing as compared to 80% missing for females, then obviously we cannot consider the data to be MCAR. If, after controlling for gender, we observe that missing data on achievement score is not related to the value of achievement score itself, then the data would be considered MAR.

Naturally the MCAR assumption is easier to test than the MAR assumption, the latter being based on the assumption that missing data on a variable is not associated in any way to the values of that variable itself. In some cases, this assumption is known to be not valid. One example is that of salary as individuals with high salaries tend to not report their salaries (Groves et al., 2004). Thus, in this case the missing data (on salary) is related to the value of missing data (with high salaries missing more often than low salaries) and the missing data mechanism is neither MCAR nor MAR.

Missing Data Handling Methods

The missing data handling methods included in this section have been individually discussed extensively in the literature spanning the last 30 years. Since any attempt to reproduce that discussion in its entirety is a complete study in its own right, only a brief description of the missing data handling methods is provided here. Readers who are interested in detailed technical aspects including mathematically-intensive proofs and theorems, and application of these methods in various fields including education are referred to Madow, Nisselson and Olkin (1983), Madow and Olkin (1983), Madow, Olkin, and Rubin (1983), Rubin (1987), Jones (1996), Groves, Dillman, Eltinge, and Little (2002), and Peugh and Enders (2004). Missing data handling methods can be divided into two broad categories. The first category includes methods that rely on discarding a portion of the sample while the second category includes methods that replace missing data with imputed values.

Case deletion methods. Two commonly used methods that work by discarding cases with incomplete information are listwise deletion and pairwise deletion.

Listwise deletion. This is one of the simplest methods of handling missing data and involves throwing away all cases that have missing data on one or more variables that are relevant to the chosen method of analysis. The remaining dataset thus consists of only those cases that have complete information. For this reason listwise deletion is also known as complete case method (McKnight et al., 2007). Although this method works well when the reduced sample is representative of the population, the smaller *n* has a direct bearing on power of the tests of hypothesis because of the negative relationship between sample size and power. In situations where the sample size obtained after listwise deletion does not represent the target population well, listwise deletion should be discarded in favor of a more appropriate missing data handling method. In SPSS, listwise deletion is the default method and is supported for a large number of procedures.

Pairwise deletion. This method is similar to listwise deletion with the difference that only cases with missing data on variables involved in a statistical procedure are removed. For example, if *X*, *Y*, and *Z* are three variables and one case in the data has a missing value on *Z*, then a procedure such as correlation will use all *n* observations to calculate r_{XY} but only $n-1$ observations to calculate r_{XZ} and r_{YZ}. This is different from listwise deletion which would have used $n-1$ cases for all three correlations. Pairwise deletion is also known as available case method (McKnight et al., 2007). Pairwise deletion has limited application in our study for two reasons. First, in models involving only one or two variables, such as one sample *t* test, independent samples *t* test, and one-way ANOVA, listwise deletion and pairwise deletion are identical methods. Second, for methods involving more than two variables, such as two-way ANOVA and multiple

regression, SPSS does not support pairwise deletion. For these reasons pairwise deletion is not considered for analytical procedures performed later in this study.

Imputation-based methods. The following methods involve replacing missing data with their imputed counterparts.

Mean imputation. This method involves replacing missing data on a variable with the mean of non-missing data for that variable. Mean imputation is also known as marginal mean imputation because the effect of other variables are not partialed out of the mean used to replace missing values (Allison, 2001). This method is one of many methods based on replacing missing data on a variable with a measure of central tendency for that variable. Depending on the measure of central tendency used, this method can take other names such as median imputation, mode imputation etc. (Chen, Jain, & Tai, 2006). Mean imputation is a faulty imputation method that is known to suppress standard error of the mean thus increasing the risk of rejecting the null hypothesis when it should not be rejected.

Regression imputation. This method involves regressing the variable with missing data on all other variables using cases that have full information for those variables. This allows computation of predicted values (or conditional means) of the variable with missing data given values of other variables. For this reason regression imputation is also known as conditional mean imputation. However, since this method does not specifically model the natural variation in missing data, it produces biased standard errors of parameter estimates (Allison, 2001).

Expectation-maximization (EM) imputation. The EM imputation is a maximum likelihood-based iterative method that involves two steps. In the first step initial values (often marginal means) are assigned to missing data. In the second step, expectations formed with those initial values are maximized. This expectation-maximization cycle is then repeated again and again until the imputed values converge based on a pre-determined convergence criteria. The EM imputation method produces unbiased standard errors of parameter estimates (Salkind & Ramussen, 2007).

Multiple imputation. Multiple imputation is an advanced imputation method that simulates the natural variation in missing data by imputing such missing data several times thus producing several complete datasets. The sets of estimates produced by these various complete datasets are then combined into a single set of estimates, for example, by averaging. Since, multiple imputation specifically models the natural variation in missing data, the standard errors of parameter estimates produced with this method remain unbiased (Rubin, 1987).

Other imputation-based methods. The four imputation methods discussed in preceding paragraphs are all supported by SPSS. There are additional imputation methods, some of which are variations of these four methods that are not directly supported by SPSS but merit mention.

Hot deck imputation. Although the hot deck method is used extensively in social science research it tends to be relatively less developed conceptually compared to other missing data imputation methods. This method involves imputing missing data on a variable for a given case by matching that case with other cases in the dataset on several

other key variables that have complete information for those cases. There are many variations of this method but one that allows for modeling of natural variability in missing data involves selecting a pool of all cases, called the donor pool, that are identical to the case with missing data (i.e. the recipient) on a number of variables and then choosing one case randomly out of that pool. The data on this randomly chosen case is then used to replace the missing value on the case with incomplete data. Another variation of the hot deck imputation method involves substituting the closest donor neighbor rather than selecting one donor from a pool of donors. This method obviously ignores the natural variability in missing data. Studies involving a large number variables require large sample sizes for hot deck imputation to work best so that cases may be matched on a number of variables. In order to select an appropriate donor, the recipient is matched with similar cases on all possible variables and not just those that are included in the method of analysis. The two requirements for selecting an external variable for use in hot deck imputation are that (1) whether external variable is associated with the variable being imputed, and (2) whether the external variable is associated with the dichotomous variable that indicates whether or not a value is missing. Other variations of the hot deck method include weighted sequential hot deck which is designed to avoid the problem of same donor being matched with a large number of recipients by restricting the number of items a donor may be chosen for imputation, and weighted random hot decks where no restriction is placed on the number of times a donor may be chosen but the donors are selected at random from the donor pool after adjusting each potential donor's probability of selection so that it matches the corresponding sample weight for that donor

(Andridge & Little, 2010). It should be noted that hot deck imputation can be used with other imputation methods such as multiple imputation where results from several imputed datasets each based on hot deck imputation are combined in order to obtain aggregate parameter estimates.

Dummy variable adjustment. This method involves constructing a separate dummy variable for each variable with incomplete data. The dummy variable is specified to take a value of 1 when the value of corresponding variable is not missing and 0 otherwise. The missing data on each variable, X is then replaced with a constant (often the marginal mean of X) and the dependent variable is regressed on all other variables including the dummy variables. The coefficient on the dummy variable corresponding to a variable X obtained in this way can be interpreted as the deviation of mean value of the dependent variable for missing data on X from the mean value of non-missing data on X. Although simple to understand and apply, this method is known to produce biased parameter estimates for the underlying method of analysis (Jones, 1996).

Zero imputation. This method simply replaces missing values on a variable with zeroes. The simplicity of this method's application is offset by its very limited usefulness. The replacement of missing data with zeroes makes conceptual sense in very specific circumstances, for example when dealing with missing achievement scores where a missing value can be reasonably assumed to occur because the respondent did not know the correct answer. However, this method produces biased parameter estimates whenever other reasons (such as anxiety or fatigue in the preceding example) are responsible for the occurrence of missing data (McKnight et al., 2007).

Single random imputation. This method can be thought of as a compromise between regression imputation and multiple imputation. The method involves regressing the variable with missing values on all other variables for cases with complete information, augmenting the resulting predicted values with random draws from the residual distribution of the regressand, and then using those augmented values to replace missing data. However, since the post-imputation dataset is treated as a complete dataset, the resulting standard errors of parameter estimates tend to be underestimates of their population counterparts (Allison, 2001).

Last observed value carried forward (LOCF). This method is commonly used in longitudinal studies and involves replacing the missing value on a variable at a certain point in time with the value of that variable from the immediately preceding time period. LOCF is known to produce biased parameter estimates with suppressed standard errors.

The imputation methods discussed in preceding paragraphs are the dominant methods in educational research. Although additional methods can be found in the current literature, most of them are variations of these methods. Since we are interested in missing data handling in context of SPSS, our focus will remain on five methods: listwise deletion, mean imputation, regression imputation, EM imputation, and multiple imputation.

Comparative Performance of Missing Data Handling Methods

Several researchers have evaluated the comparative performance of missing data handling methods during the last several decades. One of the earliest studies that compared alternative missing data handling methods was by Afifi and Elashoff (1966)

who compared listwise deletion, mean imputation, and regression imputation with simple linear regression as the methods of analysis. This study was completely analytical and did not use any simulated or empirical samples. They found that that none of the missing data handling methods that they considered was uniformly best. Their general finding was that, mean imputation works best when correlations among regression variables are low, listwise deletion works best when such correlations are moderate, and regression imputation works best when such correlations are high. In another early study, Haitovsky (1968) compared the performance of listwise deletion (also called the classical method by that author) and pairwise deletion in context of linear regression. This study simulated eight complete samples of n = 1,000 with a portion of each sample designated as missing. These eight samples differed from each other with respect to the total number of variables, the distribution of predictors, the variance-covariance matrix, and the variability in the dependent variables relative to the variability in the error term. A comparison between the regression parameter estimates obtained from the two missing data handling methods based on reduced samples with the parameter estimates of full samples revealed that listwise deletion performed best under all conditions except when the proportion of missing data was very large (in excess of .9) or when the data was missing in a very highly non-random pattern.

In the field of behavioral research Graham et al. (1996) used simulated data to evaluate missing data handling methods which included, among others, pairwise deletion, mean imputation, single random imputation, multiple imputation, and several variations of maximum likelihood imputation. Their findings suggested that under the MCAR

assumption, maximum likelihood and multiple imputation methods performed better than pairwise deletion which in turn was superior to mean imputation. However, with the exception of maximum likelihood methods, all methods developed bias in parameter estimation when the MCAR assumption was relaxed. With a sample size of 1,945, the authors used missing data percentages of 5.7% and 11.6% and their findings suggested that an increase in proportion of missing data produced larger bias in estimation. In a similar illustrative study, Wayman (2003) used 19,373 cases from a national reading test assessment that had approximately 15% missing data and four variables to compare the performance of listwise deletion, mean imputation, and multiple imputation. Based on sample means and their standard errors for normalized test scores, he concluded that multiple imputation performed the best, followed by listwise deletion, and mean imputation. However, this study did not consider effects of changes in sample size, proportion of missing data, and method of analysis.

Among the studies that looked at treatment of missing data in education, Peugh and Enders (2004) reviewed dominant missing data handling methods in educational research which they categorized into traditional and modern missing data techniques. The traditional techniques that they reviewed included listwise deletion, pairwise deletion, mean imputation, and regression imputation, while the modern techniques included maximum likelihood imputation, and multiple imputation. Based on a review of the past literature, the authors concluded that maximum likelihood and multiple imputation methods, due to their statistical properties that allow these methods to generate consistent and unbiased parameter estimates for any method of analysis, had an

almost unqualified superiority over their traditional counterparts. The authors also reviewed reporting practices related to missing data using studies from 23 applied research journals in the field of education and psychology and reported that their review of 545 studies in these journals identified 229 studies (42%) that had missing data. Of these 229 studies, only six studies reported using modern techniques such as maximum likelihood and multiple imputation while others relied on less sophisticated methods. These findings show that despite considerable publicity of known biases introduced by traditional missing data handling methods, they are frequently used by researchers in the fields of education and psychology. Peugh and Enders (2004) provide a comprehensive step by step illustration of maximum likelihood and multiple imputation methods using a longitudinal dataset which can be of value to researchers interested in implementing these methods on their own.

In a related study in education, Peng et al. (2006) compared the performance of two maximum likelihood methods (full information and expectation-maximization) and multiple imputation with listwise deletion using two real-world samples of size 1,302 and 517 in context of path analysis and logistic regression, and reported that magnitudes and/or signs of parameter estimates, *p* values in tests of hypotheses, and power can vary significantly depending on the missing data method employed. Their general conclusion was that advanced methods such as maximum likelihood imputation and multiple imputation are superior to listwise deletion when missing data is MAR. Unfortunately, the authors used samples of different sizes with different methods of analysis based on

different missing data handling methods. For this reason the interrelationships between these three factors could not be evaluated for this study.

In another recent applied study, Yesilova et al. (2011) compared several variations of the mean imputation and hot deck imputation methods under the MAR assumptions using a sample of size 4,464 and seven variables. The authors found that performance of hot deck imputation was superior to mean and median substitution methods in terms of parameters estimates, their standard errors, and correlations between real and imputed data. This study did not evaluate the effect of sample size and proportion of missing data on parameter estimates and their standard errors.

A recent review of the literature by Young et al. (2011) surveyed the performance of several missing data handling methods based on findings reported in past research. These included listwise deletion, pairwise deletion, mean imputation, mode imputation, regression imputation (both simple and multiple), hot deck imputation, expectation maximization imputation, and multiple imputation, among others. After reviewing dozens of studies, especially those from ergonomics, the authors concluded that there was no single missing data handling method that was the best in all situations. The performance of a given method was found to be dependent on factors such as proportion of missing data in the sample, sample size, distributions of variables in the sample, and the relationships between those variables. The authors concluded that even in cases where a missing data handling method worked well for a given dataset, there was no guarantee that the same method would also work well for similar datasets. Furthermore, they warned that applying a missing data handling method incorrectly without due regard

to the missing data mechanism can result in biased parameter estimates. Young et al. (2011) summarized the recommendations of various studies to provide the following guideline: when less than 1% of the data is missing, the effect of missing data handling methods is trivial; for 1-5% missing data, simple methods such as listwise deletion and regression imputation work well; for 5-15% missing data, sophisticated methods, such as multiple imputation, should be selected; and when missing data exceeds 15%, imputation results are largely meaningless regardless of the imputation method used. The authors found that there was a very limited number of studies that discussed the gain in power as a direct result of the imputation method used and recommended more research in this direction in order to allow derivation of a set of rules that can be used to select the best imputation method under a set of given conditions. Finally, the authors recommended multiple imputation as a candidate for a universal imputation method given its reliable performance even in cases where the proportion of missing data is large (up to 25%).

The Effects of Sample Size and Proportion of Missing Data on Comparative Performance of Missing Data Handling Methods

A very limited number of studies have attempted to quantify the effects of sample size and proportion of missing data on the performance of missing value imputation method. Haitovsky (1968) notes that mean imputation in linear regression can badly bias the parameter estimates. The main reason for this is because even though the overall mean does not change with mean imputation, the standard error of the mean can become considerably smaller depending on the proportion of missing data. For a variable *Y* with

n observations, *k* of which are missing but replaced by the mean of $n-k$ non-missing observations, squared standard error of the mean can be expressed as follows.

$$SE_M^2 = \frac{\sum_{i=1}^{n-k}\left(X_i - M\right)^2 + \sum_{i=n-k+1}^{n}\left(X_i - M\right)^2}{n(n-1)} \quad (1)$$

Since with mean imputation each imputed value exactly equals *M*, all deviations of imputed values from the mean are zero, i.e. $\sum_{i=n-k+1}^{n}\left(X_i - M\right)^2 = 0$, which causes SE_M^2 to become smaller. Thus, SE_M^2 is always biased as long as there is even a single imputed value that in reality deviates from the mean. It can be easily seen from the expression for SE_M^2 that it is directly proportional to the number of missing values, *k*. As *k* increases with *n* held constant (i.e. proportion of missing data increases), $\sum_{i=1}^{n-k}\left(X_i - M\right)^2$ decreases causing SE_M^2 to decrease as well. The opposite effect occurs when *k* decreases. The effect of sample size on SE_M^2 with proportion of missing data held constant is relatively less straightforward to see. Keeping *k/n* constant when *n* increases requires that *k* be increased at the same rate as *n*. Thus, a *t* times increase in both *k* and *n* would cause an increase in *n* while keeping *k/n* constant. Assuming that new observations obtained by increasing the sample size come from the same distribution, it is reasonable to expect that the deviations of such new observations from *M* are similar to those for original observations. Given this assumption, $\sum_{i=1}^{n-k}\left(X_i - M\right)^2$ increases at the rate *t* and the squared standard error of the mean for the new sample, $SE_M^{2'}$ takes the following expression.

$$SE_M^{2\,'} = \frac{t \cdot \sum_{i=1}^{n-k} \left(X_i - M \right)^2}{tn\left(tn-1 \right)} = SE_M^2 \cdot \frac{\left(n-1 \right)}{\left(tn-1 \right)} \tag{2}$$

For large *n* values, $n-1$ and $tn-1$ can be approximated by *n* and *tn* respectively, and the expression for $SE_M^{2\,'}$ simplifies to $SE_M^{2\,'} = SE_M^2 / t$, or in standard error units, $SE_M{}' = SE_M / \sqrt{t}$. In other words, the change in standard error of the mean due to an increase in sample size is inversely proportional to the square root of the rate at which that sample size increases. The tendency of standard error of the mean to become biased has been noted by other authors such as Gurland and Tripathi (1971) who have provided a correction factor when *n* is small.

For any reader who is not interested in algebraic details provided in the preceding paragraphs, the bottom-line is that whenever arithmetic mean is used to substitute for missing data, the resulting parameter estimates are biased.

Where mean imputation is one of the least mathematically sophisticated missing data imputation methods that does not take into consideration any random variability among the missing data values, multiple imputation is at the other extreme being one of the representing one of the most sophisticated imputations methods that specifically models random variation in missing data. Rubin (1987, p.114) provided a concrete formula relating the proportion of missing data to imputation efficiency for the multiple imputation method. If *m* is the number of times a complete dataset is generated and γ is the proportion of missing data, then, given *n*, relative efficiency of imputation, *E*,

measured in units of variance, can be shown to be an inverse function of proportion of missing data and a direct function of the number of imputations.

$$E = \frac{m}{m+\gamma} \tag{3}$$

When the dataset does not contain any missing data, γ takes the value of 0 and E = 1 which essentially means that no missing data is imputed and efficiency is 100%. When m is kept constant and γ increases, E decreases. For example, give $m = 1$, if γ increases from 0 to .05, E falls from 1 to .95 signifying a 5% decrease in efficiency. However, the rate of change in imputation efficiency is slower than the rate of change in the proportion of missing data. For instance, an increase in γ from 0 to .2 reduces imputation efficiency by less than four times the decrease observed for an increase in γ from 0 to .05 $\left(\Delta E = .17 \neq .20\right)$. In other words, the marginal effect of a 1% increase in γ on E decreases as γ increases. When γ is held constant, E becomes a direct function of m. For example, when γ is held constant at .05, the absolute change in E, as m changes from 1 to 2, is .02 or a 2% gain in efficiency.

The relationship between proportion of missing data and efficiency as provided by Rubin (1987) is important because it shows that for a large n, an increase in proportion of missing data can be compensated by an increase in the total number of multiple imputations. Thus, it falls to the researcher to determine how much efficiency she wants at the cost of computational complexity. The multiple imputation method itself does not impose any restrictions in this respect.

In order to clarify the effect of *m* and γ, E was calculated for selected values of *m* and γ (see Table 1). The relative efficiency calculations presented in Table 1 show that, for large samples, relative efficiency can be reasonably high $(E \geq .98)$ with just one or two imputations when proportion of missing data is low $(\gamma \leq .05)$, and with four to eight imputations when proportion of missing data is high $(.05 < \gamma \leq .20)$. These figures support the recommendation of multiple imputation as a universal imputation method of sorts as advanced by Young et al. (2011).

It should be noted that the multiple imputation method works by imputing several sets of complete datasets. Parameter estimates are then calculated from each dataset separately and the results averaged for all datasets. This is in contrast to simply averaging the values of the various datasets and calculating a single set of parameter estimates based on that averaged data. Unlike the former method, this latter approach treats the averaged dataset as a complete dataset and thus does not allow for any variation in the parameter estimates and test statistics based on those estimates. For *i* = 1, 2, ..., *m* imputations of a dataset, point estimates and their corresponding variances from each dataset, denoted by $\hat{Q}_i$ and $\hat{U}_i$, can be used to aggregate estimation results as follows. The point estimates can be averaged to obtain the aggregate point estimate.

$$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m}\hat{Q}_i \tag{4}$$

The calculation of standard error of $\bar{Q}$ requires two components, the within-imputation variance, $\bar{U}$ and the between-imputation variance, *B* which can be combined together to obtain the variance for $\bar{Q}$, *T*.

Table 1. Asymptotic Relative Efficiency of Multiple Imputation at Selected Number of Imputations (*m*) and Proportion of Missing Data (γ).

	γ				
m	0.01	0.02	0.05	0.1	0.2
1	0.99	0.98	0.95	0.91	0.83
2	1.00	0.99	0.98	0.95	0.91
3	1.00	0.99	0.98	0.97	0.94
4	1.00	1.00	0.99	0.98	0.95
5	1.00	1.00	0.99	0.98	0.96
6	1.00	1.00	0.99	0.98	0.97
7	1.00	1.00	0.99	0.99	0.97
8	1.00	1.00	0.99	0.99	0.98
9	1.00	1.00	0.99	0.99	0.98
10	1.00	1.00	1.00	0.99	0.98

These can be calculated using the following expressions.

$$\bar{U} = \frac{1}{m}\sum_{i=1}^{m}\hat{U}_i \tag{5}$$

$$B = \frac{1}{m-1}\sum_{i=1}^{m}\left(\hat{Q}_i - \bar{Q}\right)^2 \tag{6}$$

$$T = \bar{U} + \left(1 + \frac{1}{m}\right)B \tag{7}$$

The distribution of $\frac{\bar{Q}}{\sqrt{T}}$ approximately follows the *t* distribution and can be evaluated against the critical *t* value by using the following expression for degrees of freedom, *df* calculation

$$df = (m-1)\left\{1 + \frac{m\bar{U}}{(m+1)B}\right\}^2 \tag{8}$$

Although multiple imputation works very well when *n* is large, it can produce biased estimates when *n* is very small. Kim (2004) has provided the exact magnitude of this bias using Monet Carlo simulation with 50,000 samples and five imputations, for a $2 \times 3 \times 2$ factorial design. For instance, he shows that when sample size decreases from 200 to 20, the variance of the multiple imputation point estimators can increase by a factor of 10 or more when the proportion of missing data ranges between .2 and .6. This study proposed a new missing data imputation method based on multiple imputation with more desirable statistical properties than Rubin's (1987) multiple imputation method when sample sizes are very small, $(n \le 20)$.

In a comparative study Raymond and Roberts (1997) used simulation to generate multivariate data with samples of size 50, 100, and 200 comprising of four variables and applied it in a linear multiple regression context after simulating randomly missing data at 2%, 6%, and 10%. They tested several missing data handling methods such as listwise

deletion, mean imputation, and two variations of regression imputation. The authors found that, in terms of deviation from true parameter values, regression based missing data handling methods worked best while listwise deletion turned out to be the worst option. Although Raymond and Roberts (1997) considered several sample sizes, proportions of missing data, and missing data handling methods, they looked at only one method of analysis, linear multiple regression, and collapsed their findings over sample size, due to which the effect of sample size on performance of missing data handling methods could not be evaluated. Based on their analysis, these authors recommend that whenever percentage of missing data exceeds 5%, more than one missing data handling method should be used as parameter estimation results can be vary different under various methods when so much data is missing.

Alosh (2009) simulated a longitudinal count dataset to generate samples of size 30 and 60 with missing data percentages of 10% and 20% in context of a log-linear model under MCAR, MAR, and NMAR assumptions. The primary aim of this study was to compare the effect of missing data mechanisms, rather than the missing data handling methods, on parameter estimates. For this reason, this study did not focus on missing data imputation and case deletion was employed as the primary method for handling missing data. Only the MAR condition was evaluated both under case deletion and an imputation method, LOCF. The primary finding of Alosh (2009) is that under MAR and MCAR assumptions the sample estimates are very close to their true values with a maximum percent bias of about 6% while estimates obtained under NMAR assumption showed the largest biases and were clearly inferior to those obtained under MCAR and

MAR assumptions. The author also found that MAR data and MAR-LOCF data behaved similarly in context of estimation. Parameter estimates reported in this study suggested that, on average, a decrease in proportion of missing data suppressed estimation bias. On the other hand an increase in sample size increased bias, a result that seems unintuitive but may be reasonable considering the longitudinal nature of data used in this study and the fact that only two sample sizes were considered, a number that is too small to establish whether a trend is present.

A recent study that employed simulation to investigate missing data methods is Knol et al. (2010) that used an empirical sample of n = 1,338 to create 1,000 sub-samples of size 1,025 to test the performance of three missing data handling methods, listwise deletion, dummy variable adjustment (also known as missing indicator method), and multiple imputation. This study simulated missing data percentages of 2.5%, 5%, 10%, 20%, and 30% under the assumptions of MCAR and MAR and their analytical procedure, given the categorical nature of their dependent variable, involved evaluation of odds ratios. The authors found that, for their sample, the smallest deviations from true parameters were obtained with multiple imputation, then with dummy variable adjustment, with listwise deletion being the most error-prone method. This study did not consider the effect of sample size on performance of missing data handling methods and did not use any other method of analysis.

Computational Software for Missing Data Analysis

Since a large body of researchers in education rely on generally available computer packages, such as Stata, SAS, and SPSS, for their quantitative needs (Acock,

2005), it is important to address missing data handling methods in context of such packages. For this study, the focus will be on missing data handling methods that can be implemented by SPSS. The primary reason for choosing this package is that it can implement missing data imputation methods through the point-and-click interface with no need for programming. Stata and SAS on the other hand require the user to submit commands for missing data imputation using the command line interface which requires familiarity with the relevant syntax. As Acock (2005) notes: Stata is favored by those who do not have complex data management needs but who need access to cutting edge statistical procedures; SAS is preferred by researchers who need to manage complex datasets and work with this package several hours a day; and SPSS is favored by those who do not need to perform complex data management tasks or cutting edge statistical procedures and prefer a point-and-click interface. Thus, by choosing SPSS, this study is intentionally focusing on the group of researchers who are not comfortable with or in need of advanced statistical procedures and/or the command line interface. Many graduate students and researchers in education who do not specialize in quantitative research methods or some other closely related discipline fall in this group.

Although, as noted before, the current literature offers many missing data handling methods, given the difficulty involved in identification of the exact missing data process and its modeling in the estimation process when missing data is NMAR, this study will discuss missing data handling methods only under the assumptions MCAR and MAR. For MCAR data, listwise deletion works perfectly well and there is no need for missing data imputation. For MAR data, five missing data handling methods will be

investigated. These methods are listwise deletion, mean imputation, regression imputation, expectation-maximization (EM) imputation, and multiple imputation. The primary consideration for these missing data handling methods is their availability and ease of implementation in context of the target audience for this study viz. educational researchers who are interested in utilizing a range of quantitative methods available in SPSS under missing data conditions but who do not have the expertise to deploy sophisticated imputation methods that requires substantial programming with SPSS syntax.

Power Analysis

The statistical power of a test of hypothesis is the probability of rejecting the null hypothesis when the null hypothesis is actually false. It can also be described as the probability of not making a Type II error. If the probability of Type II error is denoted by β, then power equals $1-\beta$. Power is a function of three factors: Probability of Type I error, α; sample size, *n*; and the magnitude of standardized effect size desired in the population. Power is directly related to sample size and effect size, and inversely related to α. In social science applications a power of .8 is often deemed acceptable. Since there is a trade-off involved between probabilities of Type I and Type II errors, it is important to come to a practical compromise where both types of error are adequately controlled for. The usual practice in social sciences is to fix the probability Type I error at .05 and then calculate the corresponding power given sample size and effect sizes. If the magnitude of power calculated is found to be inadequate in this way, an attempt is made to increase power by either increasing the sample size or the effect size. It is not a

common practice to raise the value of α in order to gain power due to the perceived undesirability of increasing the probability of rejecting H_0 when it is in fact true. In some areas of social science research, researchers tend to put more emphasis on Type I error as compared to Type II error due to the stronger preference to avoid a false positive as compared to a false negative (Howell, 2007; Park, 2010). It should be noted that the practice of advocating large power values is controversial due to its the native relationship between probabilities of Type I and Type II error. For a related discussion, see Hoenig and Heisey (2001).

Summary

A number of past studies have compared performance of missing data handling methods in various fields including educational research. Relatively fewer studies have evaluated the effect of sample size and proportion of missing data on the performance of such methods. Unfortunately, it is not easy to synthesize the finings of these studies because they are based on different empirical samples, use different simulation techniques, and evaluate the effect of different rates of change in sample size and proportion of missing data. For this reason, as Young et al. (2011) note, it is not possible to use those findings to construct general guidelines that can help in the selection of an appropriate missing data handling method while encompassing a reasonably large subset of various possible combinations of samples size, proportion of missing data, and missing data handling method. A possible reason why such a task has not been attempted in studies targeted for publication in scientific journals is that the extent of work involved makes such a task more suitable for a book or a dissertation rather than a journal article.

A second reason is that even though the methodological awareness of missing data issues can be traced back over several decades, such awareness has only recently started to filter down to a level where some reviewers and editors specifically ask for disclosure of missing data treatment in peer-reviewed articles.

The objective of this study is to formulate general guidelines that can assist educational researchers in the selection of appropriate missing data handling methods, by using uniform empirical and simulated samples, and uniform rates of change in sample size and proportion of missing data. By keeping all of these factors constant, any observed differences in performance of missing data handling methods can more or less be attributed directly to the relative efficiency of those methods. Earlier studies, such as Roth (1994) and Young et al. (2011) have identified a need for guidelines that can help researchers choose missing data handling methods under a variety of scenarios. Furthermore, this study provides differential performance estimates for various combinations of sample size, proportion of missing data, and missing data handling method, for four separate methods of analysis that are used extensively in educational research: one sample *t* test, independent samples *t* test, two-way ANOVA, and linear multiple regression. These differential performance estimates can be used by educational researchers to apply corrections to expected bias in prior studies that involved incorrect use of missing data handling methods and consequently reported incorrect parameter estimates. It should be noted here that the four methods of analysis considered in this study are special cases of the general linear model. In this sense, the differences between these methods can be thought of as the differences in the number and types of variables

used in the general linear model, when certain factors such as sample size, proportion of missing data, and missing data handling method are all held constant. Finally, following the need for further research in this direction as identified by Young et al. (2011), this study provides an extensive power analysis to clarify how the use of missing data imputation methods can improve power of tests of hypotheses. The statistical procedures adopted to accomplish these objectives are discussed in the next section.

3. Methods

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

Data Simulation

The primary source of data used for statistical analyses performed in this study was a simulated dataset. There are two major reasons for using simulated data. First, to ensure that distributional assumptions governing the methods of analysis applied in this study were not violated. For instance, one of the methods of analysis used to test the effectiveness of missing data handling methods was two-way ANOVA. The proper implementation of this method requires that a number of underlying assumptions such as independence of observations, normal distribution of sample means, and homogeneity of dependent variable for each group, among others, be satisfied. It is difficult to satisfy all of these assumptions all of the time when two-way ANOVA is performed under a variety of experimental conditions such as different samples sizes ranging from very small to very large with data missing at different rates in these samples. The main concern is 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 is to simulate data that satisfies all underlying assumptions for analytical methods of interest and that at the

same time has characteristics that make such data suitable for analysis of real-world problems.

A second reason for using simulated data is that since we start with a complete dataset, it is relatively straightforward to observe the effect of missing data on parameter estimates by comparing results directly between complete and incomplete datasets. This allows one to determine the extent of error due to presence of missing data which in turn allows one to objectively evaluate how much of that error can be corrected by using a particular missing data imputation method.

In order to mimic data routinely encountered by educational researchers a dataset with 10,000 cases was simulated which included four continuous and two categorical variables. Examples of continuous variables frequently used in educational research include achievement scores, age, socioeconomic status, and various ability, attitude and efficacy measures. Commonly encountered categorical variables include gender, race, disability status, grade level etc. Since 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. For the four simulated continuous variables, Y, X_1, X_2, and X_3, used in this study, the variance-covariance matrix presented in Table 2 shows that Y is weakly correlated with X_1, moderately correlated with X_2, and strongly correlated with X_3. X_1, X_2, and X_3 were modeled to have significant but weak correlations with each other. This was to avoid the problem of multicollinearity in linear multiple regression models analyzed in this study.

Table 2. Variance-Covariance Matrix for the Simulated Dataset.

	Y	X_1	X_2	X_3
Y	1.0	-	-	-
X_1	0.3	1.0	-	-
X_2	0.5	0.2	1.0	-
X_3	0.7	0.2	0.2	1.0

Note. n = 10,000.

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. However, Cohen himself cautions strongly against using such rigid criteria without giving due consideration to context of the study and without considering other factors such as sample size and, probabilities of Type I and Type II errors.

Simulated Variables

Dependent variable, *Y*. This is a continuous variable, with a mean of 0 and standard deviation of 1, drawn randomly from a multivariate normal distribution of four continuous variables.

Continuous predictors, X_1, X_2, and X_3. These three predictors and *Y* have a multivariate normal distribution. For ease of interpretation all variables were specified to have a mean of 0 and standard deviation of 1.

Histograms for the dependent variable and the three continuous predictors are presented in panels *a* through *d* in Figure 1.

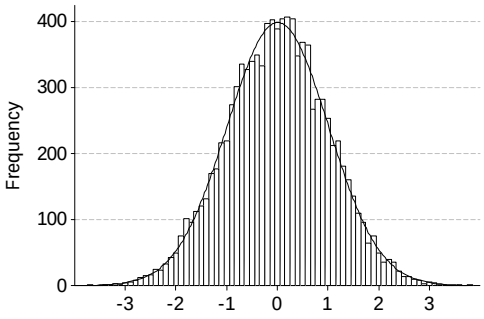
Categorical predictors, Z_1, and Z_2. Dichotomous predictor Z_1 was constructed using a uniform discrete distribution with values 0 (n = 4,945) and 1 (n = 5,055). Since the assignment of these values to Z_1 is random, this mirrors a situation where a significant mean difference in *Y* does not exist 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 MCAR assumption and not the relatively less stringent MAR assumption and as the missing values of *Y* are still independent of their own magnitude, the data is also not NMAR.

Bar charts for two categorical predictors are presented in panels *e* and *f* in Figure 1. In order to show how the means of these predictors were modeled, the marginal means and $Z_1 \times Z_2$ cell means are presented in Figure 2.

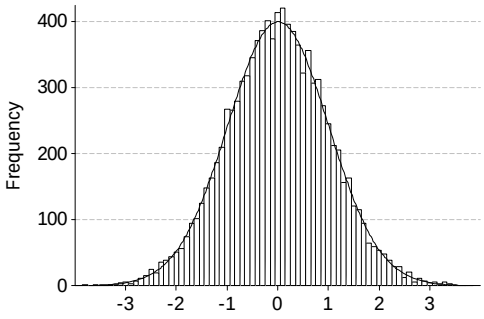
Data Analysis Approach for Simulated Data

The analyses performed with simulated data are described in the following sections.

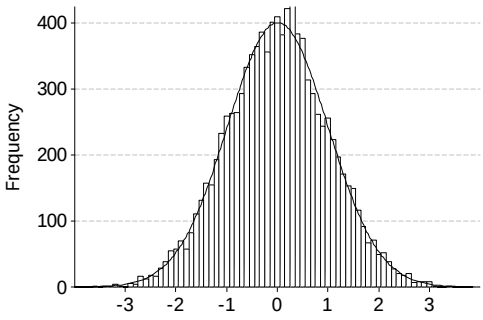
Sample selection. The simulated dataset (n = 10,000) was used to select 10 sub-samples of size 10, 20, 50, 100, 200, 500, 1000, 2000, 5000, and 10000. Each of these



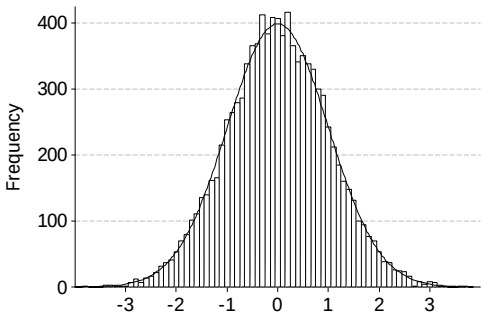
a. Y



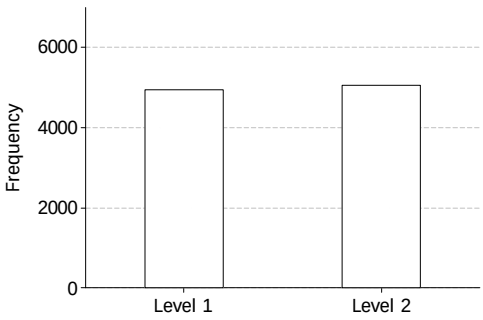
b. X_1



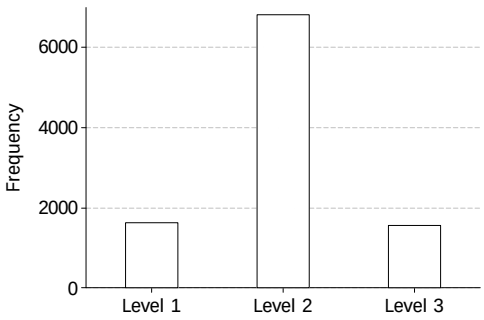
c. X_2



d. X_3

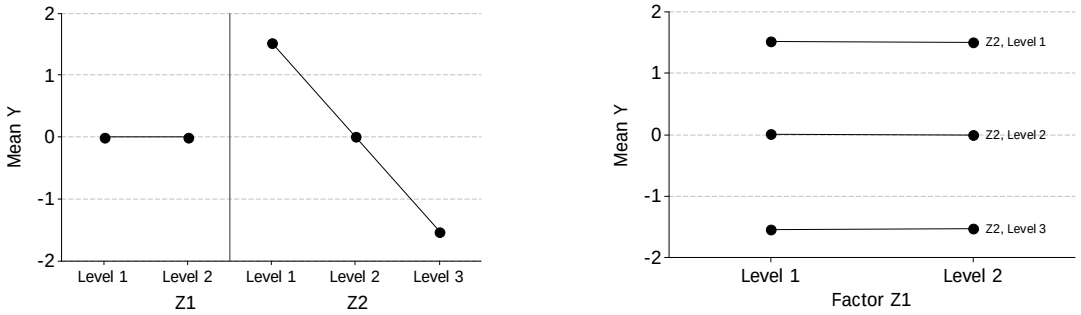


e. Z_1



f. Z_2

Figure 1. Histograms of dependent variable *Y* and the three continuous independent variables X_1, X_2, X_3, and bar charts of the two categorical variables, Z_1, and Z_2, for the complete sample, *n* = 10,000.



a. Mean *Y* at various levels of Z_1 and Z_2 b. Mean *Y* at various levels of by Z_2

Figure 2. Mean of dependent variable *Y* plotted against levels of factors Z_1 and Z_2 (a) individually and (b) as an interaction.

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. For example, five cases were randomly thrown out from a sample of size *n* = 100 in order to obtain a partial sample containing 5% missing data, *n* = 95. In order to obtain a sample with 10% missing data, ten cases were randomly removed from the original sample of *n* = 100 again rather than removing five additional cases from the *n* = 95 sample. Had the latter course been adopted, the *n* = 90, and *n* = 95 samples would not have been independent as the five observations missing from the *n* = 95 sample would have had no chance of being selected in the *n* = 90 sample. Such an

approach would have introduced inter-sample dependencies where $n = 90$ sample would have been a subset of $n = 95$ sample which would have been a subset of $n = 98$ sample and so on, rather than all samples being random draws from the complete sample of $n = 100$.

With five sub-samples for each complete sample, there was a total of 10 samples with complete data (0% missing) and 50 samples with missing data ranging between 1% and 20%. These 60 sample sizes are summarized in Table 3. A few sample sizes had to be rounded in order to avoid duplicate cell sizes. For example, in order to achieve 1% missing data from a sample size of $n = 50$, one case was removed given the obvious impossibility of removing 0.5 cases. For some very small sample sizes, some of the sample reductions were not practical. For example, for the $n = 10$ sample, a reduction of 5% cases would have meant a resulting sample size of 9.5 which rounds down to 9 and rounds up to 10. In either case, such rounding would have duplicated a pre-existing sample size. It is for this reason that there are some blank cells in Table 3.

The distribution of dependent variable, Y is shown for each sample size when there is no missing data and at each of the five missing data percentages, 1%, 2%, 5%, 10%, and 20% in Figures 3 through 12. One can observe from these figures that the distribution of Y becomes less and less normal as sample size decreases, noticeably so when the sample size is 50 or less. On the other hand, the proportion of missing data does not seem to have any effect on the distribution of Y at a given sample size. For example, in Figure 3, for a sample size of 10 we do not observe any extraordinary difference when the missing data percentage changes from 5% to 10% to 20%. The main

reason for this is that values of *Y* were drawn randomly and separately in each of the three cases depicted in Figure 3.

Table 3. Summary of Sample Sizes Used in Missing Data Analysis.

Percentage of Missing Data					
0%	1%	2%	5%	10%	20%
10	-	-	-	9	8
20	-	-	19	18	16
50	**49**	**48**	**47**	45	40
100	99	98	95	90	80
200	198	196	190	180	160
500	495	490	475	450	400
1000	990	980	950	900	800
2000	1980	1960	1900	1800	1600
5000	4950	4900	4750	4500	4000
10000	9900	9800	9500	9000	8000

Note. Sample sizes shown in boldface indicate either rounding down or adjustment by a factor of ± 1 in order to differentiate them from adjacent cells. Blank cells represent non-integer *n* values that were redundant after rounding.

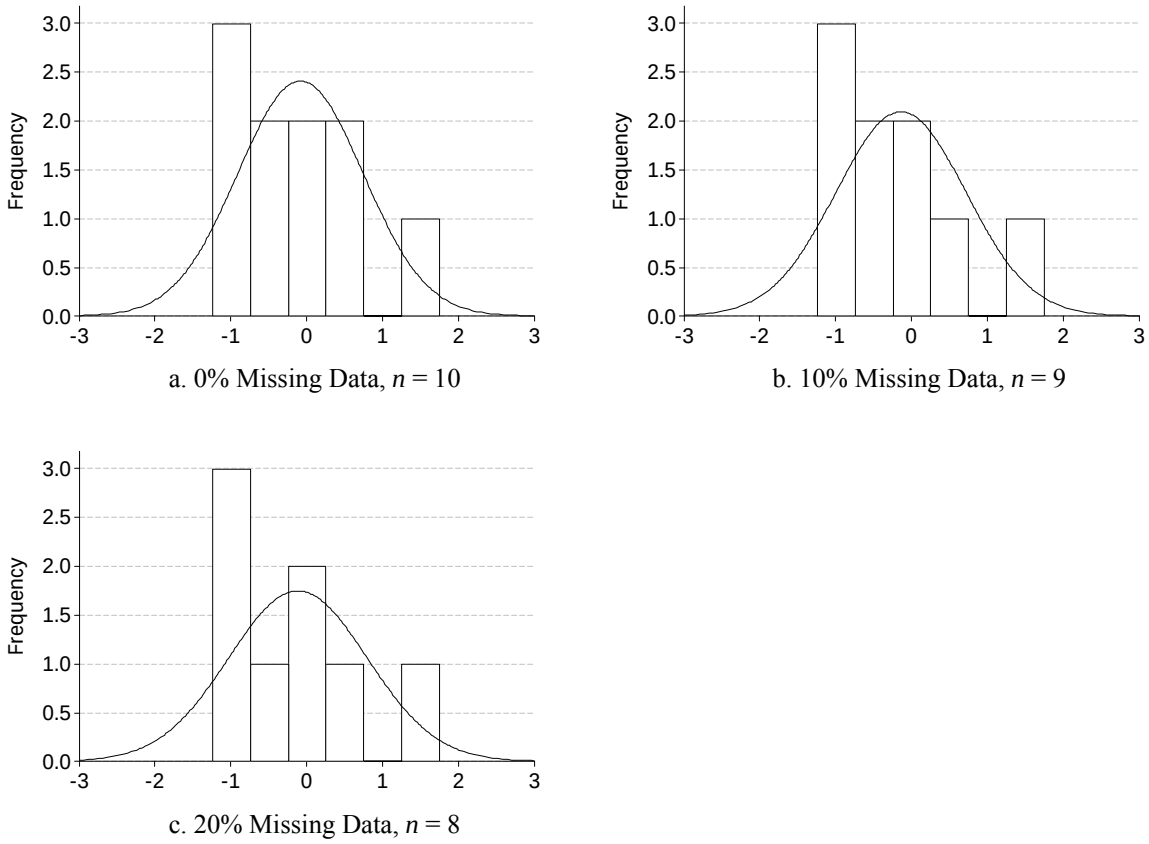
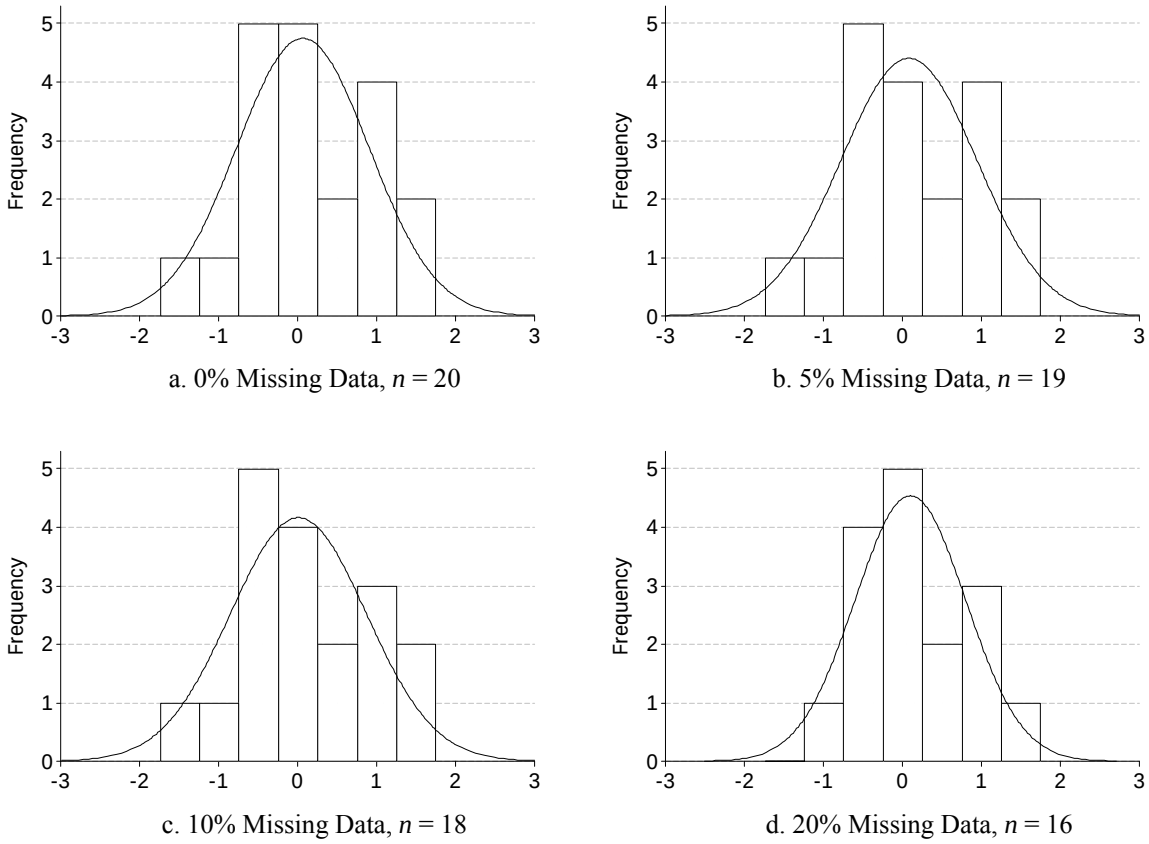


Figure 3. Histogram of the dependent variable *Y* for *n* = 10 at various percentages of missing data.



a. 0% Missing Data, $n = 20$

b. 5% Missing Data, $n = 19$

c. 10% Missing Data, $n = 18$

d. 20% Missing Data, $n = 16$

Figure 4. Histogram of the dependent variable *Y* for a sample size of 20 at various percentages of missing data.

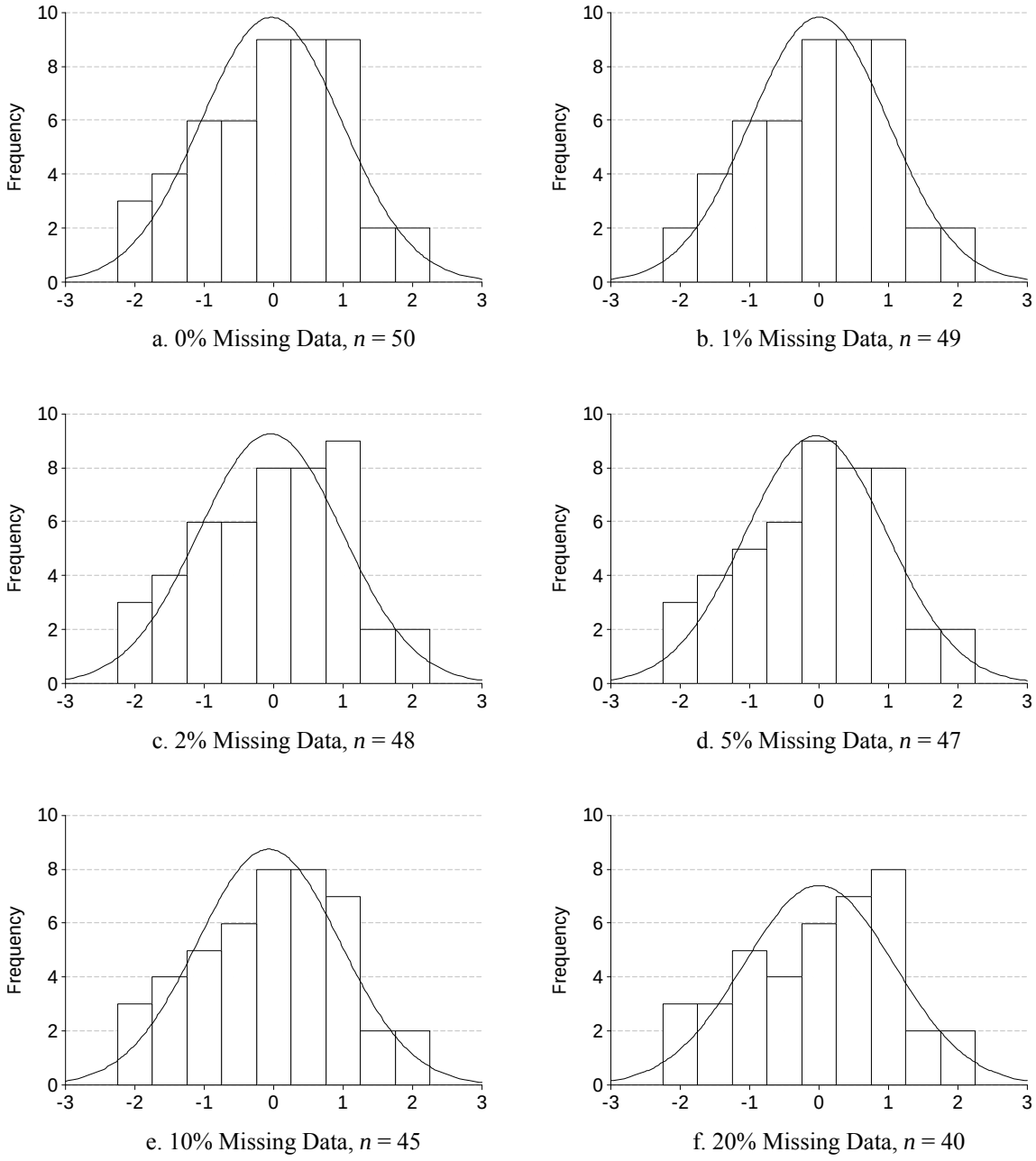
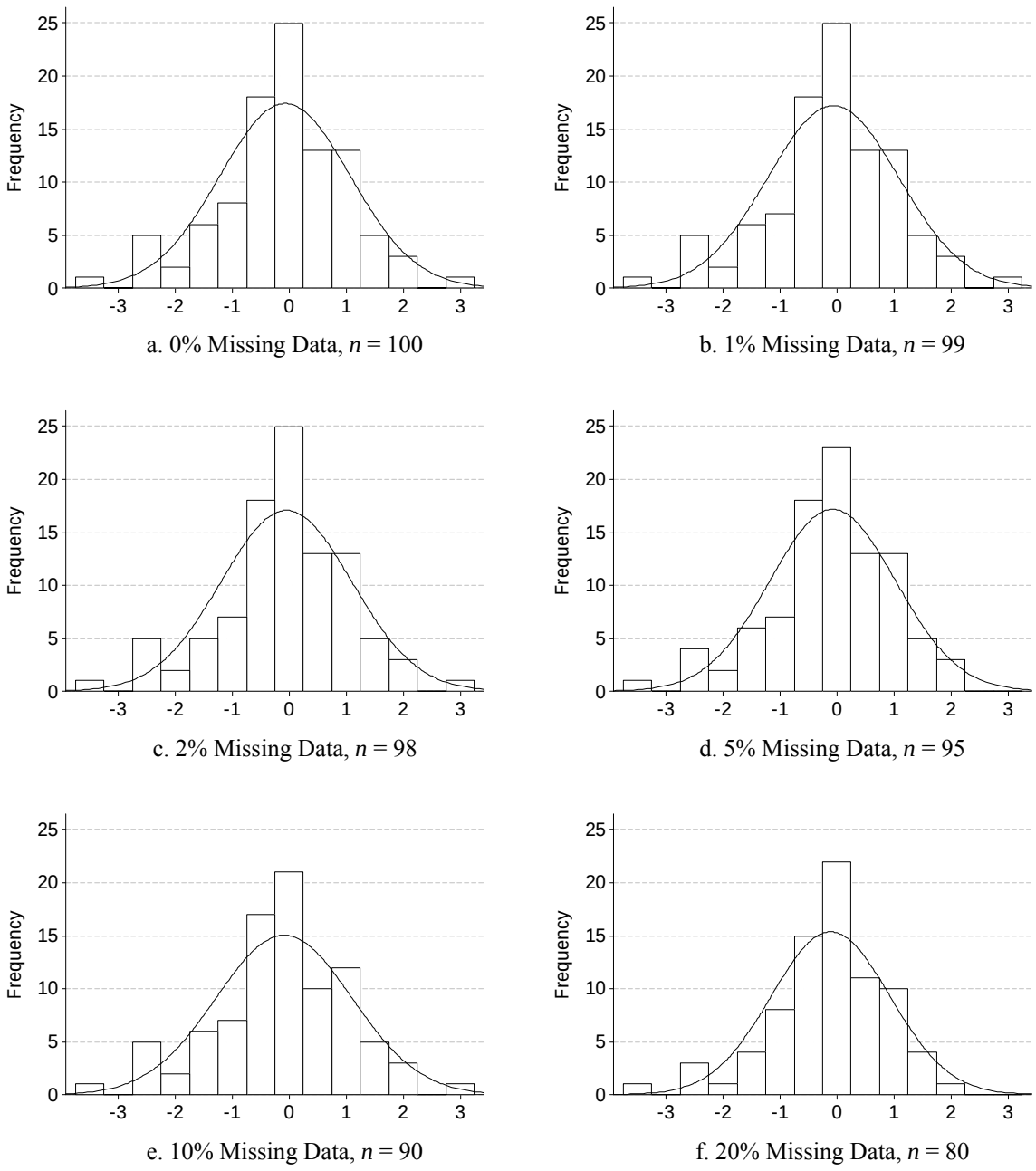


Figure 5. Histogram of the dependent variable *Y* for *n* = 50 at various percentages of missing data.



a. 0% Missing Data, $n = 100$

b. 1% Missing Data, $n = 99$

c. 2% Missing Data, $n = 98$

d. 5% Missing Data, $n = 95$

e. 10% Missing Data, $n = 90$

f. 20% Missing Data, $n = 80$

Figure 6. Histogram of the dependent variable Y for $n = 100$ at various percentages of missing data.

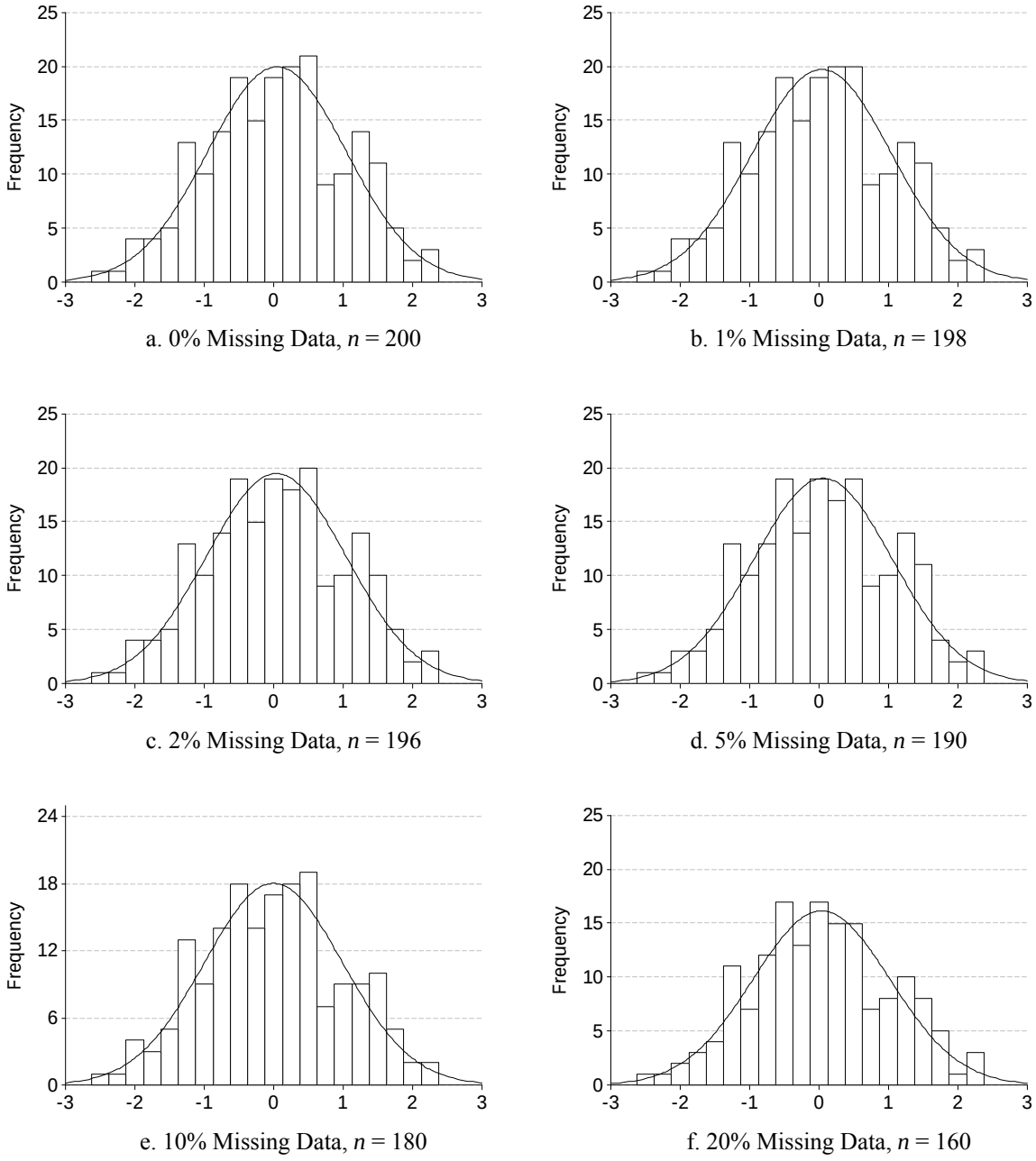
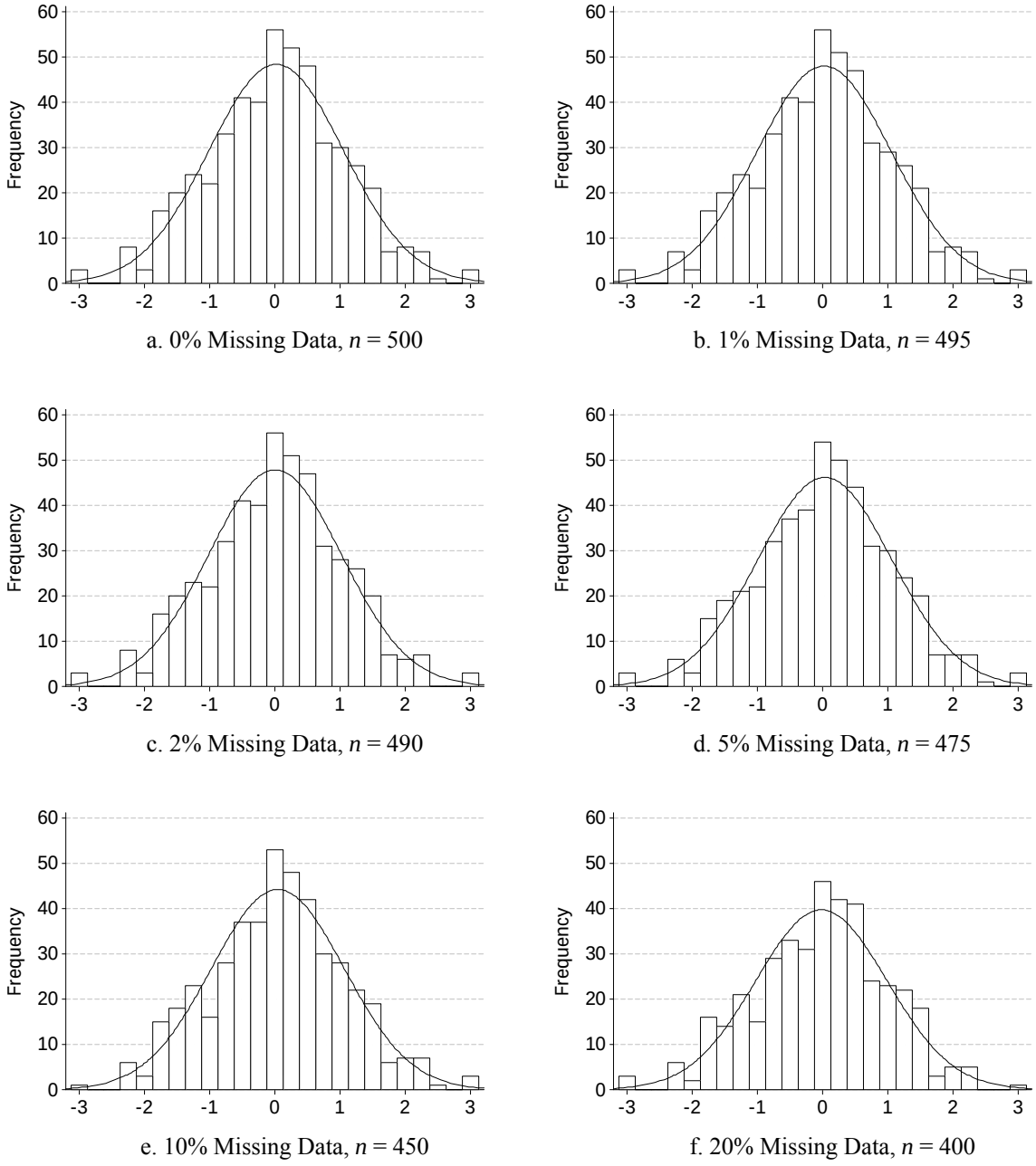


Figure 7. Histogram of the dependent variable *Y* for *n* = 200 at various percentages of missing data.



a. 0% Missing Data, $n = 500$

b. 1% Missing Data, $n = 495$

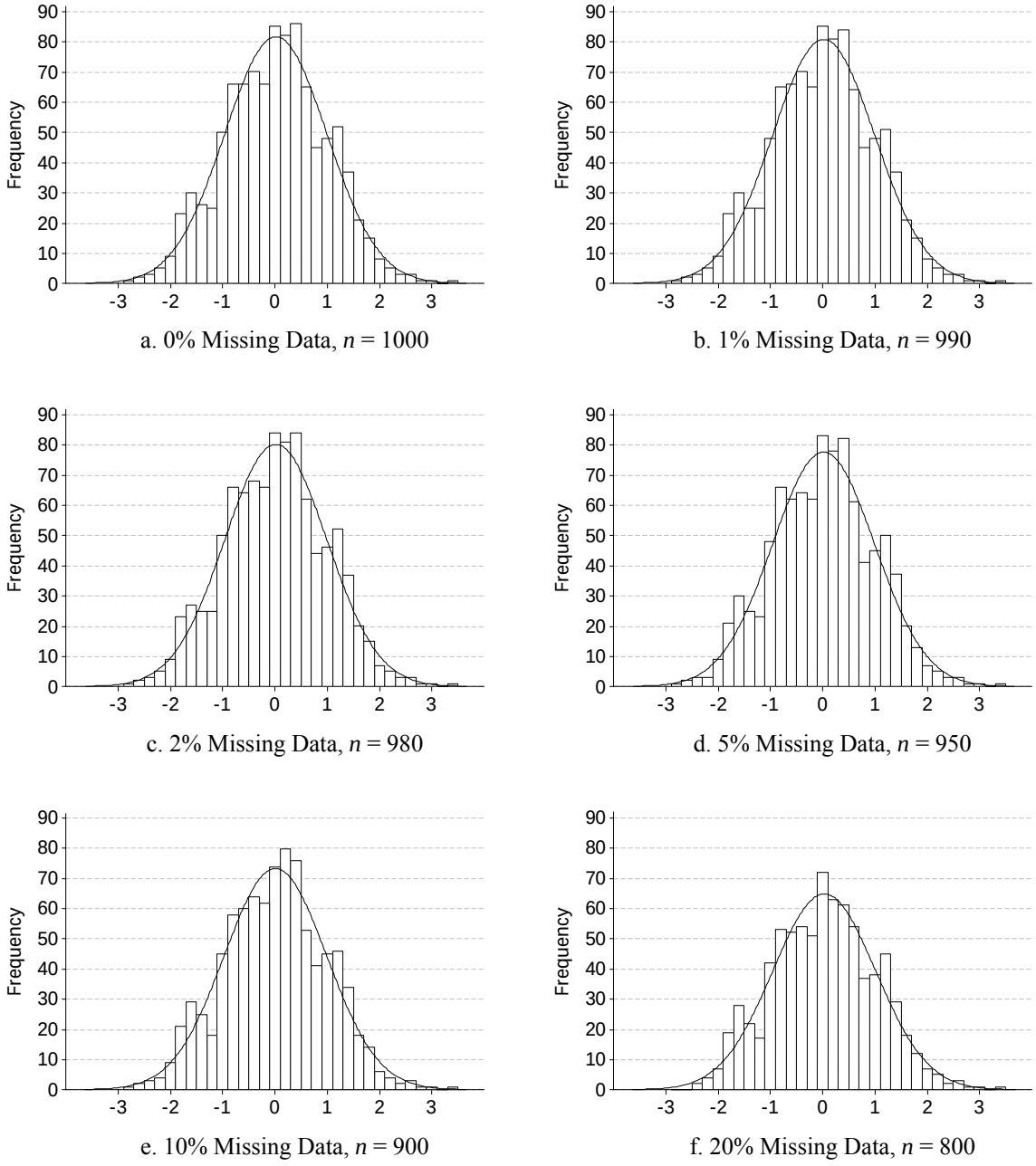
c. 2% Missing Data, $n = 490$

d. 5% Missing Data, $n = 475$

e. 10% Missing Data, $n = 450$

f. 20% Missing Data, $n = 400$

Figure 8. Histogram of the dependent variable *Y* for $n = 500$ at various percentages of missing data.



a. 0% Missing Data, n = 1000

b. 1% Missing Data, n = 990

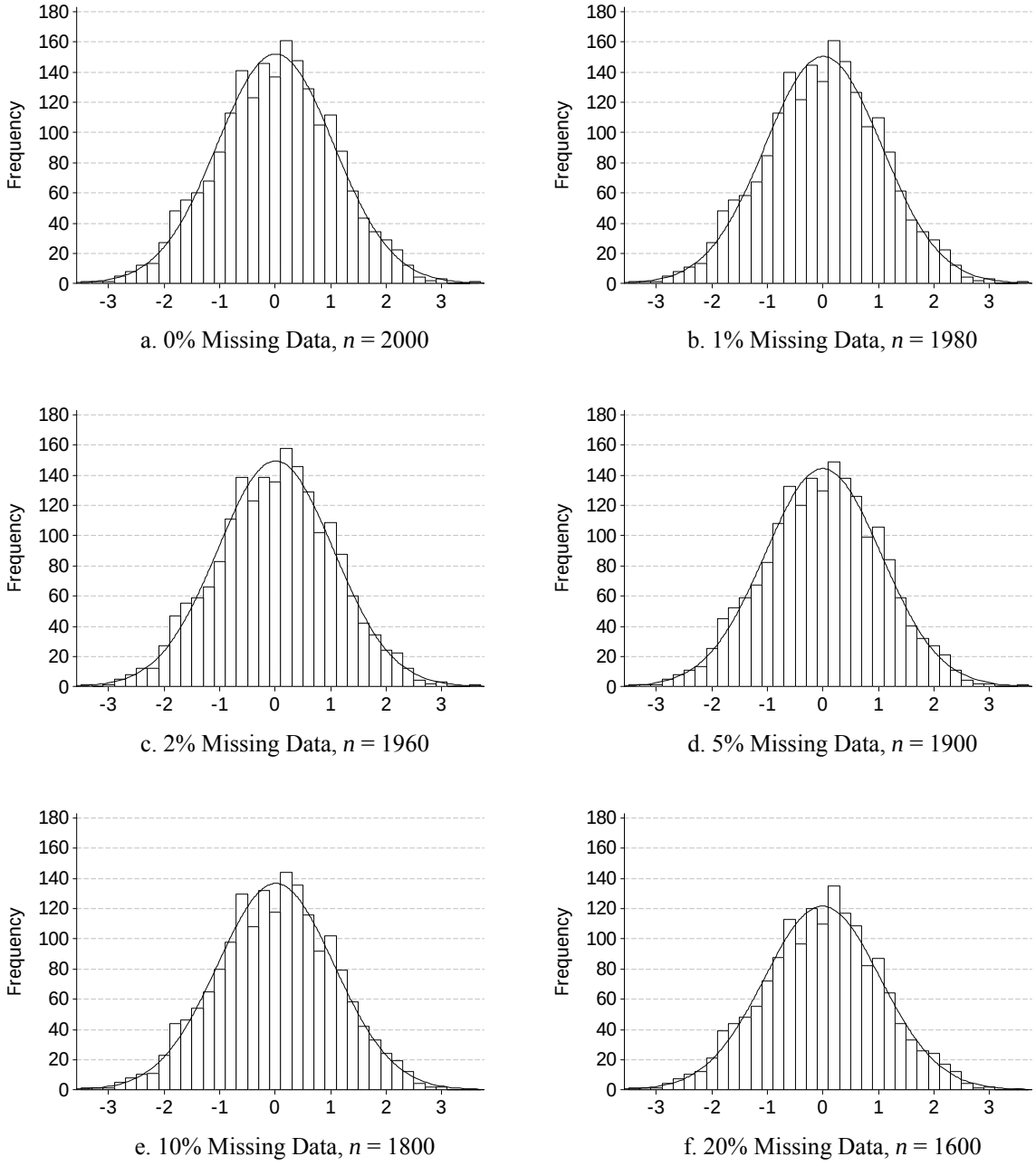
c. 2% Missing Data, n = 980

d. 5% Missing Data, n = 950

e. 10% Missing Data, n = 900

f. 20% Missing Data, n = 800

Figure 9. Histogram of the dependent variable Y for n = 1000 at various percentages of missing data.



a. 0% Missing Data, $n = 2000$

b. 1% Missing Data, $n = 1980$

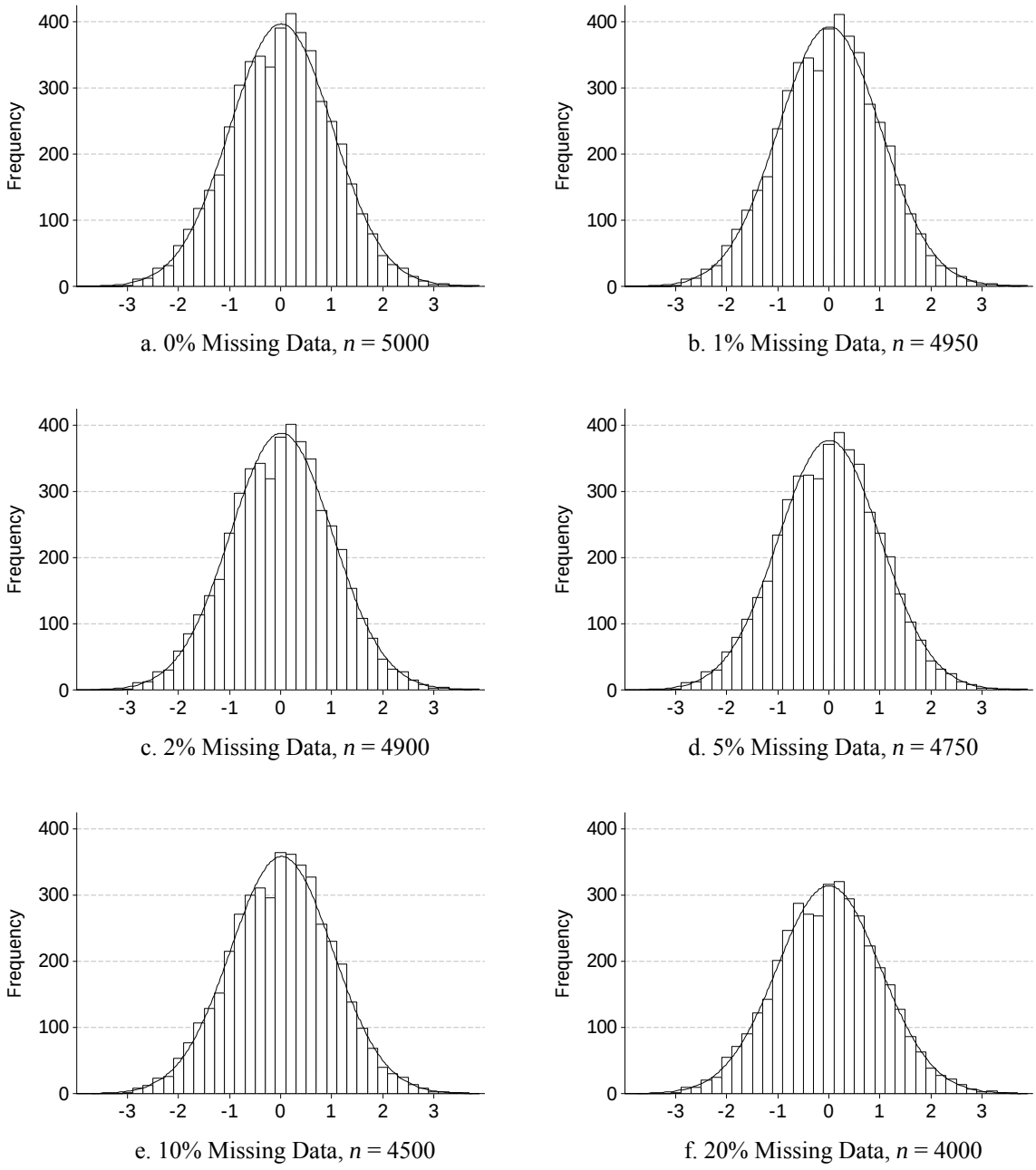
c. 2% Missing Data, $n = 1960$

d. 5% Missing Data, $n = 1900$

e. 10% Missing Data, $n = 1800$

f. 20% Missing Data, $n = 1600$

Figure 10. Histogram of the dependent variable Y for $n = 2000$ at various percentages of missing data.



a. 0% Missing Data, $n = 5000$

b. 1% Missing Data, $n = 4950$

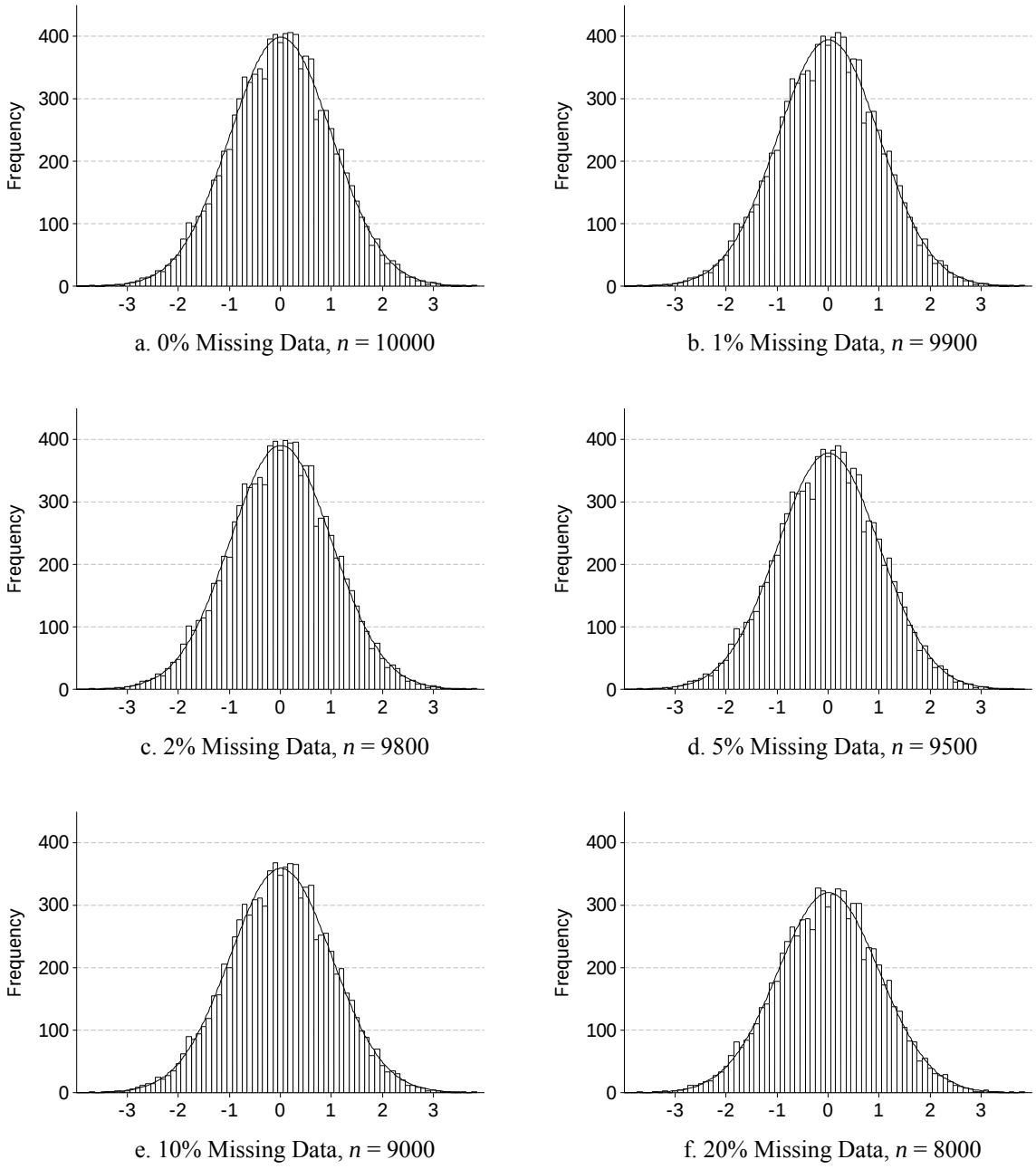
c. 2% Missing Data, $n = 4900$

d. 5% Missing Data, $n = 4750$

e. 10% Missing Data, $n = 4500$

f. 20% Missing Data, $n = 4000$

Figure 11. Histogram of the dependent variable Y for $n = 5000$ at various percentages of missing data.



a. 0% Missing Data, n = 10000

b. 1% Missing Data, n = 9900

c. 2% Missing Data, n = 9800

d. 5% Missing Data, n = 9500

e. 10% Missing Data, n = 9000

f. 20% Missing Data, n = 8000

Figure 12. Histogram of the dependent variable Y for n = 10000 at various percentages of missing data.

Methods of analysis. 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 were 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 ranging from simple models with just one variable (for example, one sample *t* test) to advanced models with a mix of continuous and categorical variables (for example, multiple regression).

The model specification for each of the four methods of analysis depended upon the number of type of variables available in the simulated dataset and thus in this respect mimics the real-world decision-making process where a researcher identifies her research questions first, then operationalizes her variables and finally chooses an appropriate method based on the number and type of those variables. The nature of variables simulated in this study did not leave much choice for the model specifications used in the four methods of analysis. For the one sample *t* test the only dependent variable, *Y* was utilized. 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. There are of course some other model specifications that could have been used. For example, the independent samples *t*

test could have been used with two of the three levels of Z_2. Similarly, multiple regression could have used two dummy variables to represent Z_2 as an additional predictor of *Y*. However, the important thing to note is that the focus here is not on exact functional form of the variables. As a matter of fact the exact functional form does not even matter for the simulated variables because their inter-relationships are not based on any underlying review of the literature. The primary focus of simulated data analysis is on how the various missing data handling methods perform under various methods of analysis. Since the four methods of analysis selected for this study are already a subset of hundreds of methods of analysis available in educational research, if different functional forms for the same method are also treated as distinct methods of analysis, such an approach is unlikely to have a significant effect on the findings of this study.

Missing data handling methods. Five missing data handling methods that are available in SPSS and that can be readily implemented with the point-and click interface without any need to write an elaborate syntax were selected for missing data analysis. These methods are listwise deletion, mean imputation, regression imputation, EM imputation, and multiple imputation. It is stressed once again that the primary consideration for these missing data handling methods is their availability and ease of implementation in context of the target audience for this study viz. educational researchers who are interested in utilizing a range of quantitative methods available in SPSS under missing data conditions but who do not have the expertise to deploy sophisticated imputation methods that often require substantial programming with SPSS syntax. 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.

Model comparisons. For each of the four methods of analysis, model parameter estimates and associated tests of hypotheses were separately 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 directly 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. It is important to note that this model comparison approach does not involve the critical values of the test statistic. This is so because for the models based on simulated data we do not really care about whether the model is statistically significant or not. We would be interested in that aspect of data analysis in a latter part of this study when dealing with real-world data and where models are based on empirical inter-variable relationships supported by underlying educational research.

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. The normalized root mean squared error for m observations of $X_{Observed}$ that have a true value of X_{True} can be expressed as follows.

$$Normalized\ RMSE = \frac{\sqrt{\frac{\sum_{i=1}^{m}\left(X_{Observed,i} - X_{True}\right)^2}{m}}}{Max(X) - Min(X)} \tag{9}$$

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. In our study, the F and t statistics come from models based on different

methods of analysis and can only be compared with their counterparts within the same methods of analysis and not across methods of analysis. For example, errors in *F* statistic computed under multiple regression for two different imputed samples can be compared with each other but not with the error in *F* statistic observed for two-way ANOVA models because the two methods of analysis are based on different variables and thus are not measuring identical changes in the dependent variable. Normalized *RMSE* resolves this issue by standardizing observed error values allowing comparisons across methods of analysis. A disadvantage of using normalized RMSE is that it is affected by any extreme values in the data. A fix for this problem is to use the median to calculate a root median squared error, *RMdSE*, a measure that ignores extreme values. However, as a median-based measure *RMdSE* has limited usefulness as (1) it is not based on all data points, and (2) it lends itself poorly to algebraic manipulations. Another error measure that can be used in place of RMSE is mean absolute error (*MAE*) which is simply the average of absolute deviations of individual values from their true counterparts. *MAE* can be calculated using the following expression.

$$MAE = \frac{\sum_{i=1}^{m} \left| X_{Observed,i} - X_{True} \right|}{m} \tag{10}$$

The main advantage of using *MAE* is that it has the same units as *X* and is thus easy to interpret *MAE* is also straightforward to calculate. A disadvantage of this measure is that it is not very useful for comparing quantities measured using different units of measurement. For instance, in our statistical analysis, we cannot use *MAE* to compare *t* values obtained from a one-sample *t* test with an *F* value obtained from a two-way

ANOVA because *t* and *F* statistics have different units. In this respect normalized RMSE is superior to MAE as the former measure is unit free and thus can be used to measure variables measured in different units of measurement.

Power analysis. Power was computed separately under each method of analysis for each of the original 10 samples containing no missing or imputed data (see Table 4). It should be noted that the power values supplied in this table are relevant only to original datasets with no missing data and to datasets that become complete after missing data imputation. They are not relevant to samples based on listwise deletion because listwise deleted sample sizes are smaller and we know that sample size has a direct bearing on power (McKnight et al., 2007). Power calculations for samples based on listwise deletion are presented separately for each method of analysis elsewhere in this study. The reason for presenting power analysis for full sample sizes is to give the reader an *a priori* idea about how much of Type II error we are controlling for. For example, for a one-sample *t* tests, power of .8 or more is achieved at a sample size of 50 (exact power = .93) while for independent samples *t* tests, a similar level of power is attained at a sample size of 200 (exact power = .94). This difference in sample sizes required to provide the same magnitude of power for two different methods of analysis is due to the difference in number and types (categorical or continuous) of variables used in those methods.

The power calculations presented in Table 4 are based on a medium effect size as defined by Cohen (1992): for a one-sample *t* test and independent samples *t* test the measure of effect size is *d* which is the standardized difference between two means; for two-way ANOVA the measure of effect size is *f* which is the standardized treatment

Table 4. Power of the Test to Detect Medium Effect Size for Various Analysis Methods.

	Analysis Method			
n	One Sample *t* Test	Independent Samples *t* Test	Two factor ANOVA	Multiple Regression
10	.293	.105	.061	.088
20	.565	.184	.092	.188
50	.933	.410	.207	.524
100	.999	.697	.426	.874
200	1.000	.940	.775	.997
500	1.000	1.000	.996	1.000
1000	1.000	1.000	1.000	1.000
2000	1.000	1.000	1.000	1.000
5000	1.000	1.000	1.000	1.000
10000	1.000	1.000	1.000	1.000

Note. Medium effect size is defined as Cohen's $d = .5$ for one sample and independent samples *t* tests, $f = .25$ for two-way ANOVA, and $f^2 = .15$ for multiple regression Power calculations assume two-tailed *t* tests, six groups for two-way ANOVA, one set of four predictors in multiple regression, and $\alpha = .05$.

standard deviation; and for multiple regression the measure of effect size is f^2 which is simply the proportion of explained to unexplained variation in the dependent variable. Cohen (1992) suggests $d = .5$, $f = .25$, and $f^2 = .15$ for medium effect sizes. Power analysis conducted in G*Power revealed that in order to obtain a power of .8 at 5% level of significance, the minimum sample size required is $n = 34$ for a two-tailed one sample *t*

test, $n = 128$ for a two-tailed independent samples t test ($n = 64$ per group), $n = 211$ for two-way ANOVA with six groups, and $n = 85$ for linear multiple regression with one set of four predictors. For each method of analysis, the sample sizes presented in Table 4 cover the entire range of possibilities from being much smaller than to being much larger than the *a priori* sample sizes required to achieve a power of .8 .

Empirical Data Example 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. Empirical data comes from the Program for International Student Assessment which is an international assessment of literacy in mathematics, reading, and science of 15-year old students. In the U.S., National Center for Education Statistics (NCES) is responsible for all domestic administrational aspects of this assessment. The PISA survey is administered to the 15-year old U.S. high school student population every three years with one of the three subjects, mathematics, reading, and science being the focus in any given survey year. In the 2003 administration of PISA the subject in focus was mathematics and the U.S. sample size was 5,456 (NCES, 2003). The questionnaire for this survey is 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 is to test the effectiveness of guidelines constructed on the basis of simulated data.

Empirical measures. The following variables were selected for data analysis from the PISA dataset. The selection was based on similarity of characteristics of these variables with their simulated counterparts.

Dependent variable. The dependent variable is math achievement which is distributed normally, measured on a continuous scale, and ranges between 200 and 800. PISA reports math achievement as a set of plausible values that represent random draws from the set of all possible scores that can be attributed to a student (OECD, 2005). Although the best way to evaluate this type of data is to repeat the analysis five times, once with each plausible value, and then combine the five sets of results (Wu, 2005), such an exercise is beyond the scope of this study as we are primarily interested in assessing the accuracy of estimates with and without missing data which can only be determined when the complete dataset is known. This is obviously a situation in which there is no need to make inferences about unknown population parameters specifically because those parameters are not unknown. In our case for instance, the objective is to select a complete set of values on variables of interest from the PISA dataset, designate some of the data as missing, impute those missing values, and then compare estimation results for a model of interest between complete, incomplete, and imputed datasets in order to evaluate the merits of using missing data handling methods. Although it is certainly possible to estimate population parameters after imputing missing data in an incomplete dataset, such an exercise does not allow the calculation of exact error between true parameter values and their estimates without which we cannot evaluate the effectiveness of missing data handling methods. For simplicity, the five plausible values for each student are simply averaged into a single score. Such averaging can also be justified by pointing to the high inter-plausible value correlations which were all in

excess of .90 . Readers who are interested in detailed technical discussions of plausible values are referred to Mislevy (1991), and Mislevy, Beaton, Kaplan, and Sheehan (1992).

Continuous predictors. 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. The construction of reading achievement is similar to that of the dependent variable as it is also reported as a set of five plausible values. Like math achievement, these five plausible values have high correlations among themselves, the smallest coefficient of correlation being .85 . Reading achievement is normally distributed and ranges between 200 and 800.

Math anxiety. This variable is a measure of anxiety felt by a student when engaged in math-related tasks. Math anxiety is measured on a continuum, normally distributed, and standardized to have a mean of 0 and standard deviation of 1.

Home educational resources. This variable is a measure of 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 standardized to have a mean of 0 and standard deviation of 1.

A comparison between the variance-covariance matrices of simulated and empirical predictors (see Table 2 and 5) shows that there are some differences between them and that some of the correlation coefficients for empirical variables have negative signs. However, what is more important to note is the similarity in the pattern of

relationship among the four variables which show that math achievement is correlated somewhat weakly with home educational resources, moderately with math anxiety, and strongly with reading achievement. This pattern is not very different from that observed between *Y* and its three continuous predictors. Similarly, the inter-predictor correlations presented in Table 5 are also weak like their simulated counterparts. 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.

Table 5. Variance-Covariance Matrix for the Empirical Dataset.

	Math achievement	Home Educational resources	Math anxiety	Reading achievement
Math achievement	1.0	-	-	-
Home educational resources	0.3	1.0	-	-
Math Anxiety	-0.4	-0.1	1.0	-
Reading achievement	0.8	0.3	-0.3	1.0

Note. n = 5,456

Categorical predictors. Two categorical predictors with the same number of categories as Z_1 and Z_2 were selected from the PISA 2003 dataset.

Gender. This variable has two categories: male, n = 2,740; and female, n = 2,715. One case had a missing value for gender.

Grade. This variable has three categories: grade 9, n = 1,667; grade 10, n = 3,339, and other, n = 448. Two cases had a missing value for grade.

For the dependent variable, and continuous and categorical predictors discussed in preceding paragraphs, the PISA 2003 dataset had three missing values, one for gender and two for grade reducing the maximum number of observations available for analysis from 5,456 to 5,453. It should be noted here that since PISA data is being used for illustrative purposes in this study, there is not much value in generating the missing data for the actual sample (n = 5,456). The values of continuous predictors reported by PISA 2003 are actually based on individual item responses that are mapped onto a continuum based on principles of item response theory (IRT). A detailed discussion of IRT methods and their application is beyond the scope of this study. Interested readers should consult relevant textbooks (Baker, 2001; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Lord, 1980) and the PISA 2003 technical report (OECD, 2005) for detailed technical information including functional forms of the IRT models and their underlying assumptions. 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. We can certainly use a dataset that contains missing data to begin with and then impute such data. However, such an exercise will not allow us to compare the effectiveness of missing data handling methods and thus any findings from such an exercise will have very limited usefulness.

Proposed empirical data analysis approach. The empirical variables described in preceding paragraphs were used to evaluate the effectiveness of missing data handling

guidelines supported by simulated data. In order to accomplish this objective the U.S. portion of PISA 2003 sample was used to construct a dataset that had no missing data and that comprised of the four continuous and two categorical variables described earlier. As it has been shown already, these variables have characteristics that are very similar to those of the simulated variables.

Once the empirical dataset was selected, a portion of it was designated as missing. The dataset containing missing data was then analyzed using the same missing data handling methods that were employed in 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 any of the accuracy lost due to missing data could be recovered with the use of missing data handling methods. Since the main steps involved in the proposed empirical data analysis are not different from those followed for simulated data analysis, the methods and techniques that were used for simulated data analysis were replicated, albeit at a relatively smaller scale, for empirical data analysis.

An important point to note here is that the PISA data comes with sampling weights that are required to make the PISA sample representative of the target population. Although parameter estimates can be obtained with or without use of sampling weights in SPSS for any method of analysis, only estimates that take proper weights into consideration are representative of the population. Unweighted results obtained from the sample cannot be generalized to the population. Since the issue of sampling weights is important only at the time of application of method of analysis and

has no bearing on the missing data handling method (which occurs before the method of analysis is applied), for the purposes of this study, either of the two approaches, weighted or unweighted, is appropriate. In order to keep the discussion simple and to avoid diverging from the main focus of this study, the missing data analysis for PISA 2003 dataset will not employ sampling weights. Readers interested in the effect of ignoring sampling weights on parameter estimates are referred to Hahs-Vaughn (2005).

Empirical Data Example 2

In order to evaluate the effectiveness of missing data handling methods for smaller datasets, a smaller empirical dataset was employed. This data comes from the Population and Housing portion of decennial U.S. Census published by the U.S. Census Bureau (2000). Although the U.S. Census covers the entire United States and its territories, 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 in example 1, the objective of this example 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 example 1 where the intention is to use a relatively advanced method of analysis viz.

multiple regression, empirical example 2 employed a simpler method viz. independent samples *t* test to ensure a broader coverage of analytical methods chosen for this study.

Computational Software Considerations

IBM SPSS Statistics 18 was used for all analyses performed in this study with the exception of simulated data generation which was done with Minitab 16 and power analyses which were done with G*Power 3. For various corporate administrative reasons earlier versions of IBM SPSS Statistics have been known as SPSS, SPSS Statistics and PASW Statistics. Since all of these names refer to essentially the same product, in order to avoid confusion, this program is always referred to as SPSS in this study. It is also important to note the specific SPSS version used because earlier versions of this program did not support some imputation methods. For example, multiple imputation method became available only in version 18 of this product.

The G*Power program was used for all power computations presented in this study. This program allows computation of one of the following parameters, effect size, power, probability of Type I error, or sample size when values of the remaining three parameters are provided as input. The program allows power calculations for a number of analytical methods ranging from simple, such as one sample *t* test, to advanced, such as repeated measures MANOVA. For a detailed description of all analytical methods supported by this program, the formulas used for power computations, and information about program availability see Faul, Erdfelder, Lang, and Buchner (2007).

4. Results

Simulated Data

Results of analytical procedures described in the methods 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 are presented in Tables 6, 7, 8, and 9. All of these correlations are significantly different from zero at 5% level of significance and show a general decreasing trend in magnitude from left to right as the percentage of missing data increases. Furthermore, the correlations tend to be stronger for expectation maximization imputation and multiple imputation methods as compared to mean imputation and regression imputation methods. When proportion of missing data is 5% or less, almost without exception, all imputation methods produce correlations between original and imputed data that are in excess of .95. Only for sample sizes that are less than 50 with percentage of missing data exceeding 5% do we see somewhat weaker correlations, in one case falling as low as .74. Mean imputation seems to work well as long as the percentage of missing data is 10% or less but the correlation between mean imputed and original data falls quickly regardless of sample size as this percentage exceeds 10%. It should be noted that there is no table of correlations for listwise deletion. The reason for this is that since coefficient of

Table 6. Coefficient of Correlation Between Actual and Imputed Data with Mean Imputation.

	Percentage of Missing Data				
n	1%	2%	5%	10%	20%
10	-	-	-	.98***	.97***
20	-	-	.99***	.97***	.74***
50	.97***	.99***	.97***	.96***	.95***
100	.99***	.99***	.94***	.99***	.81***
200	.99***	.99***	.97***	.94***	.88***
500	.99***	.98***	.97***	.93***	.87***
1000	.99***	.99***	.98***	.95***	.90***
2000	.99***	.99***	.97***	.95***	.89***
5000	.99***	.99***	.97***	.95***	.90***
10000	.99***	.99***	.98***	.95***	.89***
Mean	.99	.99	.97	.96	.88

Note. For the entire table, $\bar{r} = .95$.

*$p < .05$; **$p < .01$; ***$p < .001$

correlation can only be calculated for pairs of observations, with listwise deletion correlation is calculated only for pairwise valid observations and since the values in each pair are exactly identical, the correlation coefficient is always 1. 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, EM imputation, and multiple imputation were .95, .96, .98, and .98, suggesting that the such correlation is

Table 7. Coefficient of Correlation Between Actual and Imputed Data with Regression Imputation.

	Percentage of Missing Data				
n	1%	2%	5%	10%	20%
10	-	-	-	.78**	.84**
20	-	-	.99***	.97***	.86***
50	.99***	.97***	.96***	.92***	.95***
100	.99***	.99***	.99***	.97***	.94***
200	.99***	.99***	.99***	.97***	.93***
500	.99***	.98***	.97***	.97***	.94***
1000	.99***	.99***	.98***	.96***	.93***
2000	.99***	.99***	.98***	.96***	.92***
5000	.99***	.99***	.98***	.96***	.93***
10000	.99***	.99***	.98***	.97***	.93***
Mean	.99	.99	.98	.94	.92

Note. For the entire table, $\bar{r} = .96$.

*$p < .05$; **$p < .01$; ***$p < .001$

strongest for EM 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.

Before we look at the relative performance of various missing data handling methods, it is important to note that the arithmetic mean and its standard error plays a central role in all four analytical methods considered in this study. For this reason it is important to evaluate the effect of various missing data handling methods on these

Table 8. Coefficient of Correlation Between Actual and Imputed Data with Expectation Maximization Imputation.

	Percentage of Missing Data				
n	1%	2%	5%	10%	20%
10	-	-	-	.92***	.85**
20	-	-	.99***	.97***	.86***
50	.99***	.99***	.99***	.99***	.95***
100	.99***	.99***	.99***	.98***	.97***
200	.99***	.99***	.99***	.98***	.96***
500	.99***	.99***	.99***	.98***	.96***
1000	.99***	.99***	.99***	.99***	.97***
2000	.99***	.99***	.99***	.98***	.96***
5000	.99***	.99***	.99***	.98***	.97***
10000	.99***	.99***	.99***	.98***	.96***
Mean	.99	.99	.99	.98	.94

Note. For the entire table, $\bar{r} = .98$.

*$p < .05$; **$p < .01$; ***$p < .001$

important statistics. For one sample *t* test with listwise deletion and mean imputation, the sample mean of *Y* and its relative error are presented in Table 10. The reason for presenting means from the two missing data handling methods in the same table is that in this particular case the two sets of values were completely identical. The reason for this is that the sample mean of *Y* does not change when missing *Y* values are replaced by that mean. The figures in Table 10 show two general results. First, when proportion of missing data is held constant, the sample mean value of *Y* approaches the population

Table 9. Coefficient of Correlation Between Actual and Imputed Data with Multiple Imputation.

	Percentage of Missing Data				
n	1%	2%	5%	10%	20%
10	-	-	-	.99***	.95***
20	-	-	.99***	.96***	.75**
50	.99***	.99***	.99***	.98***	.93***
100	.99***	.99***	.99***	.99***	.96***
200	.99***	.99***	.99***	.98***	.97***
500	.99***	.99***	.99***	.98***	.95***
1000	.99***	.99***	.99***	.98***	.96***
2000	.99***	.99***	.99***	.98***	.97***
5000	.99***	.99***	.99***	.98***	.96***
10000	.99***	.99***	.99***	.98***	.96***
Mean	.99	.99	.99	.98	.94

Note. For the entire table, $\bar{r} = .98$.

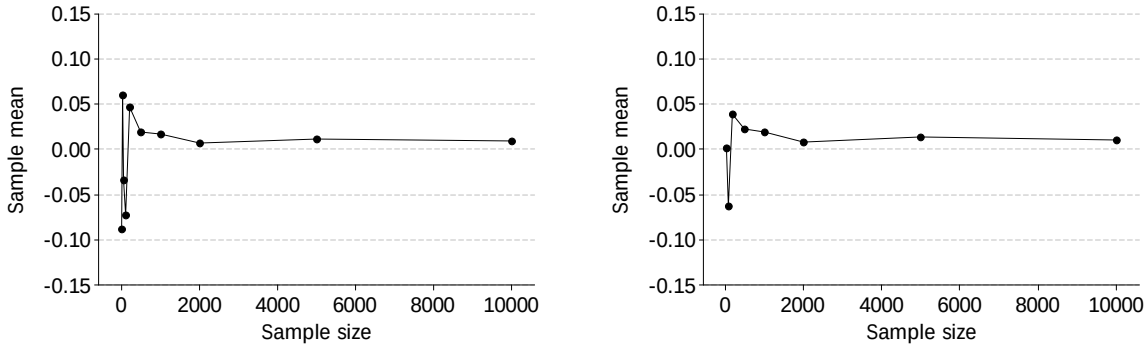
*$p < .05$; **$p < .01$; ***$p < .001$

mean, $\mu_Y = 0$. This can be seen more clearly in Figure 13 where mean *Y* is plotted as a function of sample size separately for various proportions of missing data. The second result that can be observed in Table 10 is that when sample size is held constant, the effect of increase in proportion of missing data on the mean of *Y* is not very clear. This can be seen in Figure 14 where sample mean of *Y* is plotted as a function of proportion of missing data at various sample sizes. Although the straight line connecting the means falls closer and closer to the $Y = 0$ line as sample size increases, there is no clear upward

Table 10. Sample Mean and its Relative Error under One Sample *t* Test with Listwise Deletion or Mean Imputation.

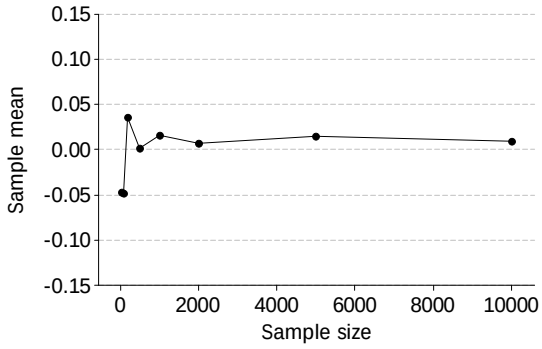
		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
M	10	-0.088	-	-	-	-0.144	-0.121
	20	0.060	-	-	0.077	0.004	0.095
	50	-0.034	0.001	-0.047	-0.042	-0.070	-0.002
	100	-0.072	-0.063	-0.049	-0.075	-0.091	-0.111
	200	0.048	0.040	0.036	0.059	-0.034	0.035
	500	0.019	0.022	0.002	0.033	0.044	-0.019
	1000	0.017	0.019	0.017	0.018	0.009	0.030
	2000	0.007	0.008	0.007	0.001	0.012	-0.009
	5000	0.012	0.014	0.015	0.010	0.021	-0.001
	10000	0.010	0.011	0.009	0.010	0.011	0.011
	Mean	-0.002	0.007	-0.001	0.010	-0.024	-0.009
Relative Error of *M*	10	-	-	-	-	0.643	0.385
	20	-	-	-	0.268	-0.937	0.565
	50	-	-1.038	0.375	0.230	1.068	-0.956
	100	-	-0.134	-0.329	0.030	0.256	0.537
	200	-	-0.166	-0.235	0.235	-1.723	-0.263
	500	-	0.150	-0.917	0.705	1.301	-2.005
	1000	-	0.098	-0.052	0.017	-0.489	0.747
	2000	-	0.191	-0.015	-0.794	0.706	-2.294
	5000	-	0.149	0.248	-0.157	0.702	-1.050
	10000	-	0.071	-0.071	-0.031	0.133	0.102
	Mean	-	-0.085	-0.125	0.056	0.166	-0.423

Note. Mean relative error of M = -0.083; *RMSE* = 0.024 .

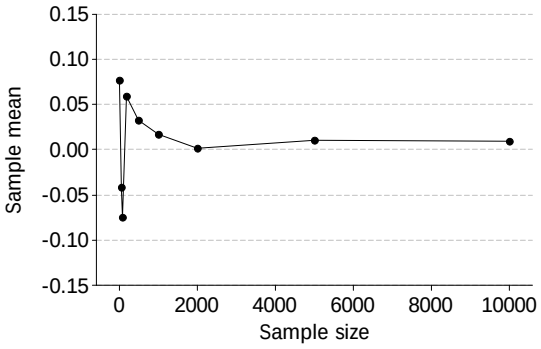


a. 0% Missing Data

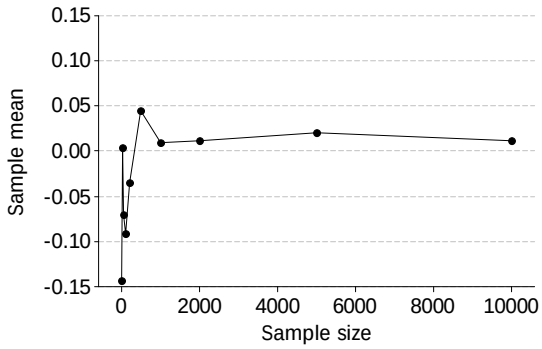
b. 1% Missing Data



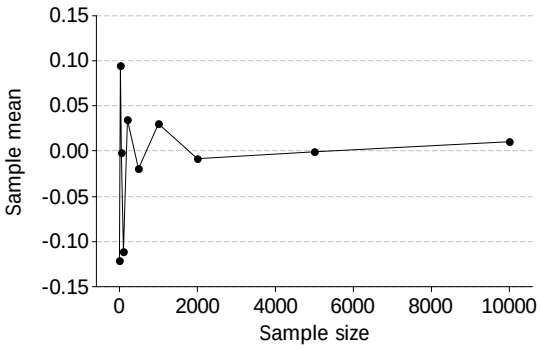
c. 2% Missing Data



d. 5% Missing Data

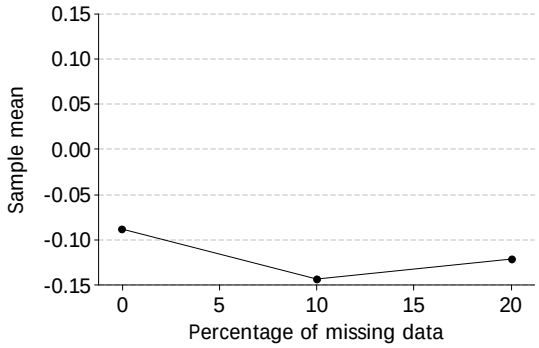


e. 10% Missing Data

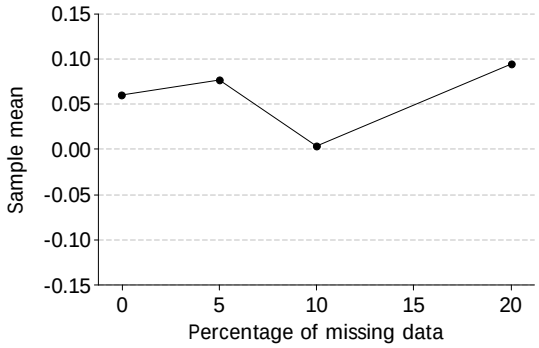


f. 20% Missing Data

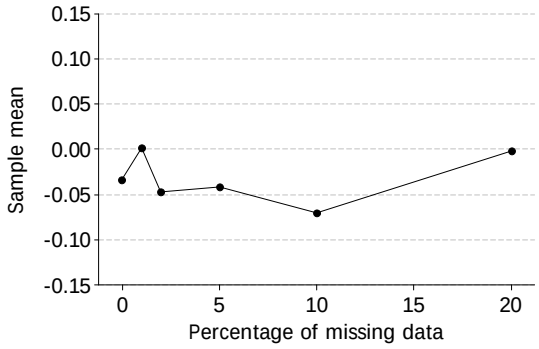
Figure 13. Sample mean plotted as a function of sample size for missing sample data ranging between 0% and 20% under one sample *t* test with listwise deletion and mean imputation.



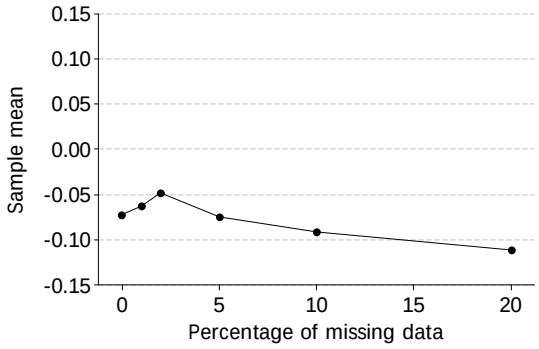
a. $n = 10$



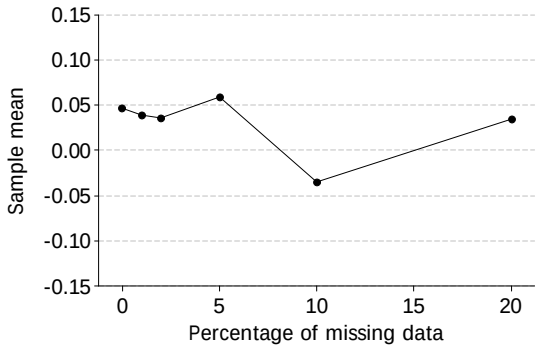
b. $n = 20$



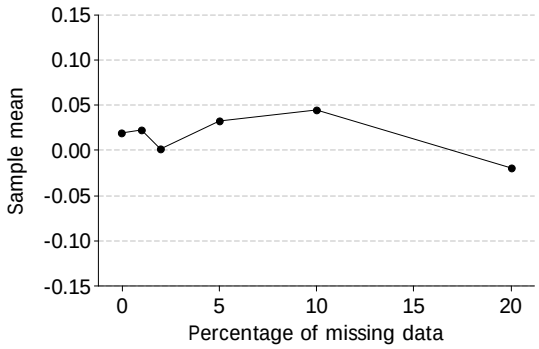
c. $n = 50$



d. $n = 100$



e. $n = 200$



f. $n = 500$

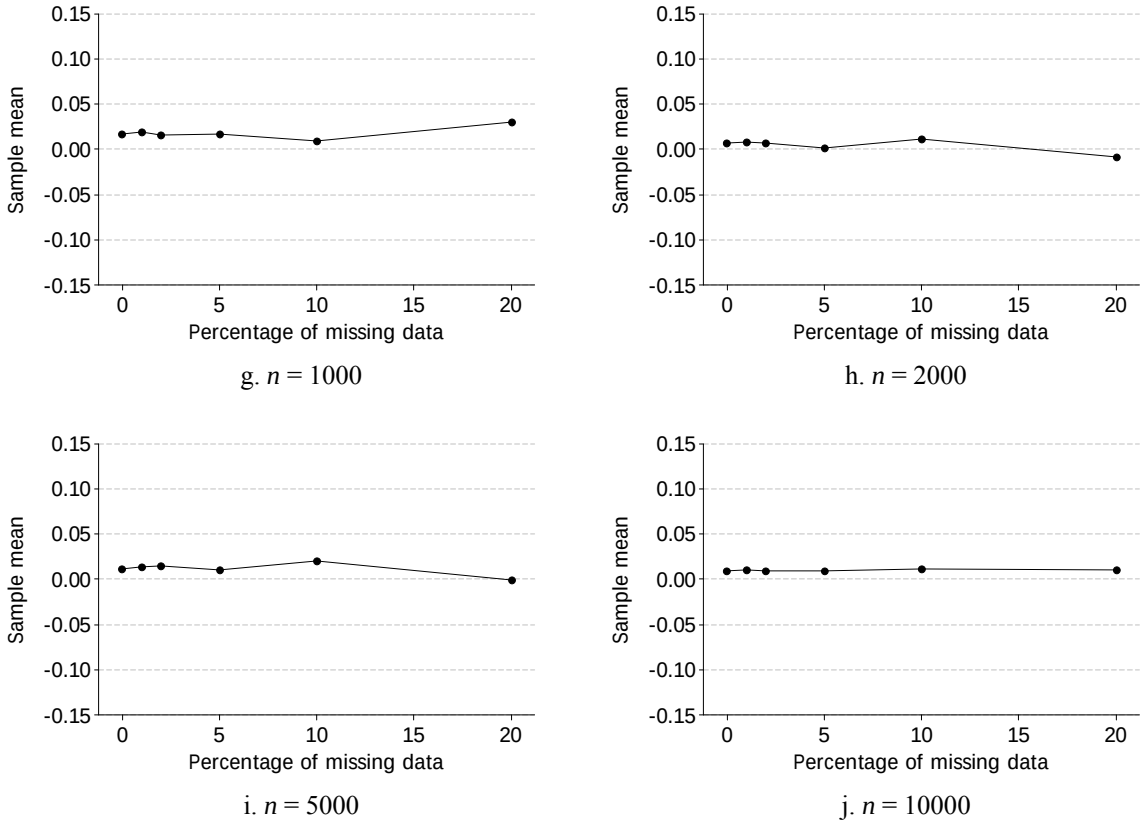
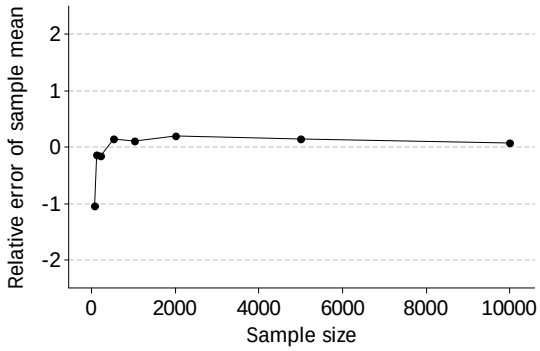


Figure 14. Sample mean plotted as a function of percentage of missing data for sample size ranging between 10 and 10,000 under one sample *t* test with listwise deletion and mean imputation.

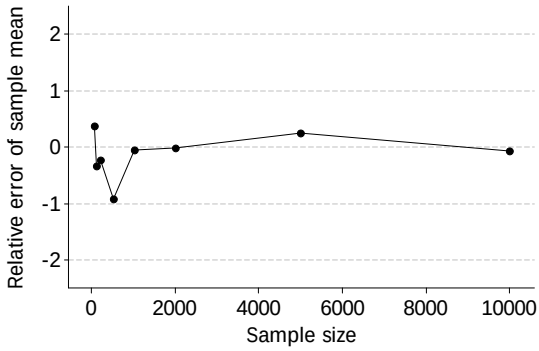
or downward trend for the ten sample sizes in general. The implication here is that when data is missing at random, the effect of proportion of missing data on sample mean is also random with listwise deletion and mean imputation. However, with these two missing data methods, the effect of sample size on sample mean is not random because the sample mean becomes more and more precise as *n* increases.

The relative error of the sample mean is plotted as a function of sample size in Figure 15, and as a function of proportion of missing data in Figure 16. This relative error is simply the difference in sample mean for a reduced sample and its complete counterpart expressed as a percentage of the latter. The graphs presented in Figure 16 show that, holding proportion of missing data constant, with listwise deletion or mean imputation the relative error decreases as sample size *n* increases i.e. sample mean becomes more and more precise. Figure 16 on the other hand shows that, holding sample size constant, the relative error becomes negligible only when the sample size is very large (n = 10,000). Even at a sample size of 5,000 (panel *i*) the fluctuations in relative error of the mean can be large.

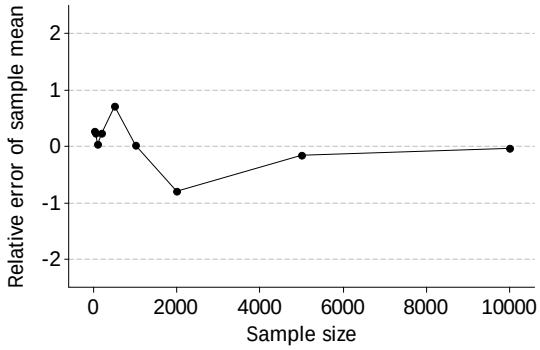
For one sample *t* test, standard error (*SE*) of the mean of *Y* and relative error of that *SE* are presented in Table 11 for listwise deletion and in Table 12 for mean imputation. It is important to deal with the standard error of the mean because the *t* statistic which is used to test the significance of mean in one sample *t* test, is by itself nothing but the ratio of deviation of sample mean from the test value, to the standard error of that mean. It can be observed from the figures presented in Tables 11 and 12 that as sample size increases, the value of SE decreases. This can be seen very clearly in Figure 17 which plots standard error of the mean as a function of sample size at various proportions of missing data. Although the standard errors are not identical under listwise deletion and mean imputation, the differences are very small. What is important to note is that a very clear decreasing trend is present in all panels of Figure 17. This means that



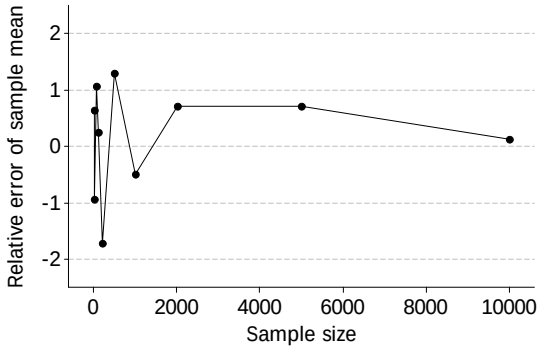
a. 1% Missing Data



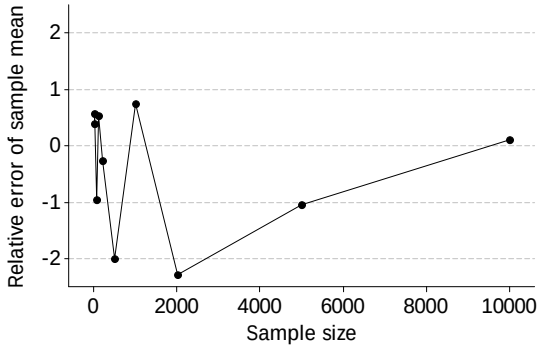
b. 2% Missing Data



c. 5% Missing Data

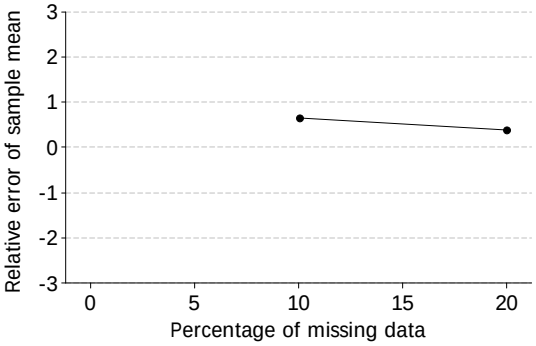


d. 10% Missing Data

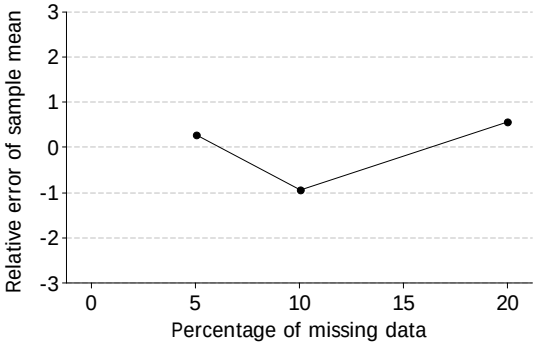


e. 20% Missing Data

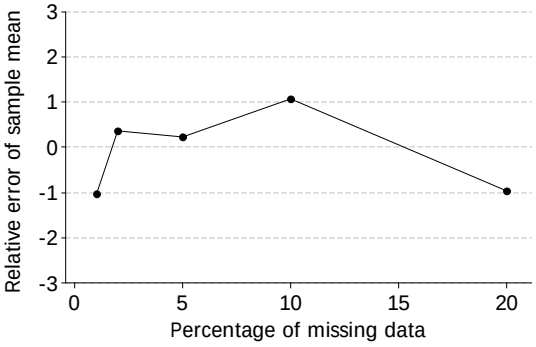
Figure 15. Relative error of sample mean plotted as a function of sample size for missing sample data ranging between 0% and 20% under one sample *t* test with listwise deletion and mean imputation.



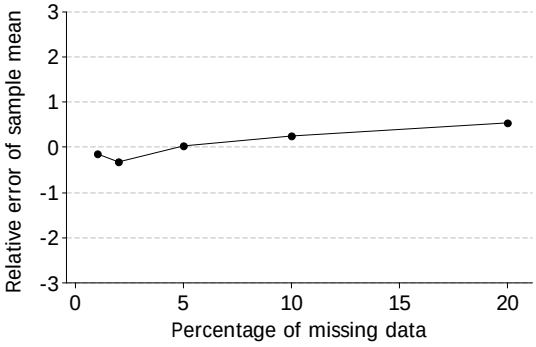
a. $n = 10$



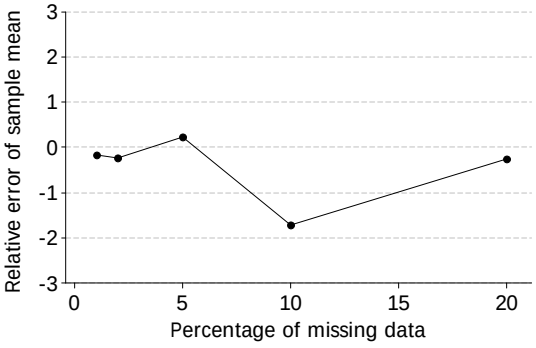
b. $n = 20$



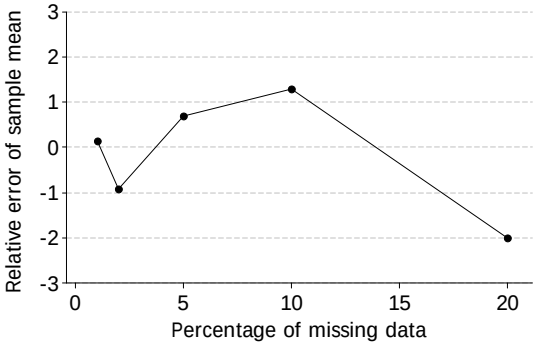
c. $n = 50$



d. $n = 100$



e. $n = 200$



f. $n = 500$

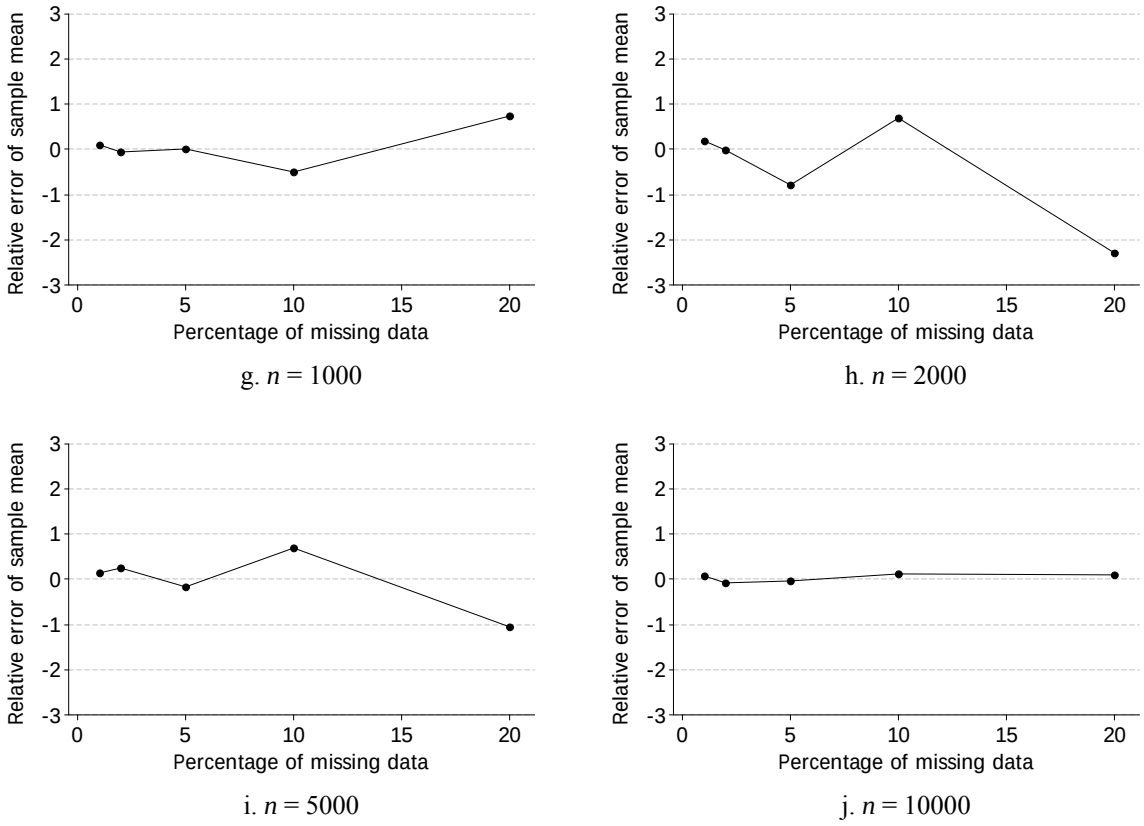


Figure 16. Relative error of sample mean plotted as a function of percentage of missing data for sample size ranging between 10 and 10,000 under one sample *t* test with listwise deletion and mean imputation.

as sample size increases, observed value of the *t* statistic also increases (because standard errors becomes smaller) and thus it becomes easier to reject the null hypothesis (when in fact this should not happen).

In order to easily visualize the difference in standard errors between the two missing data handling methods, *SE* is plotted against the proportion of missing data separately for various sample sizes in Figure 18. The individual graphs in Figure 18

Table 11. Standard Error of the Mean and its Relative Error under One Sample *t* Test with Listwise Deletion.

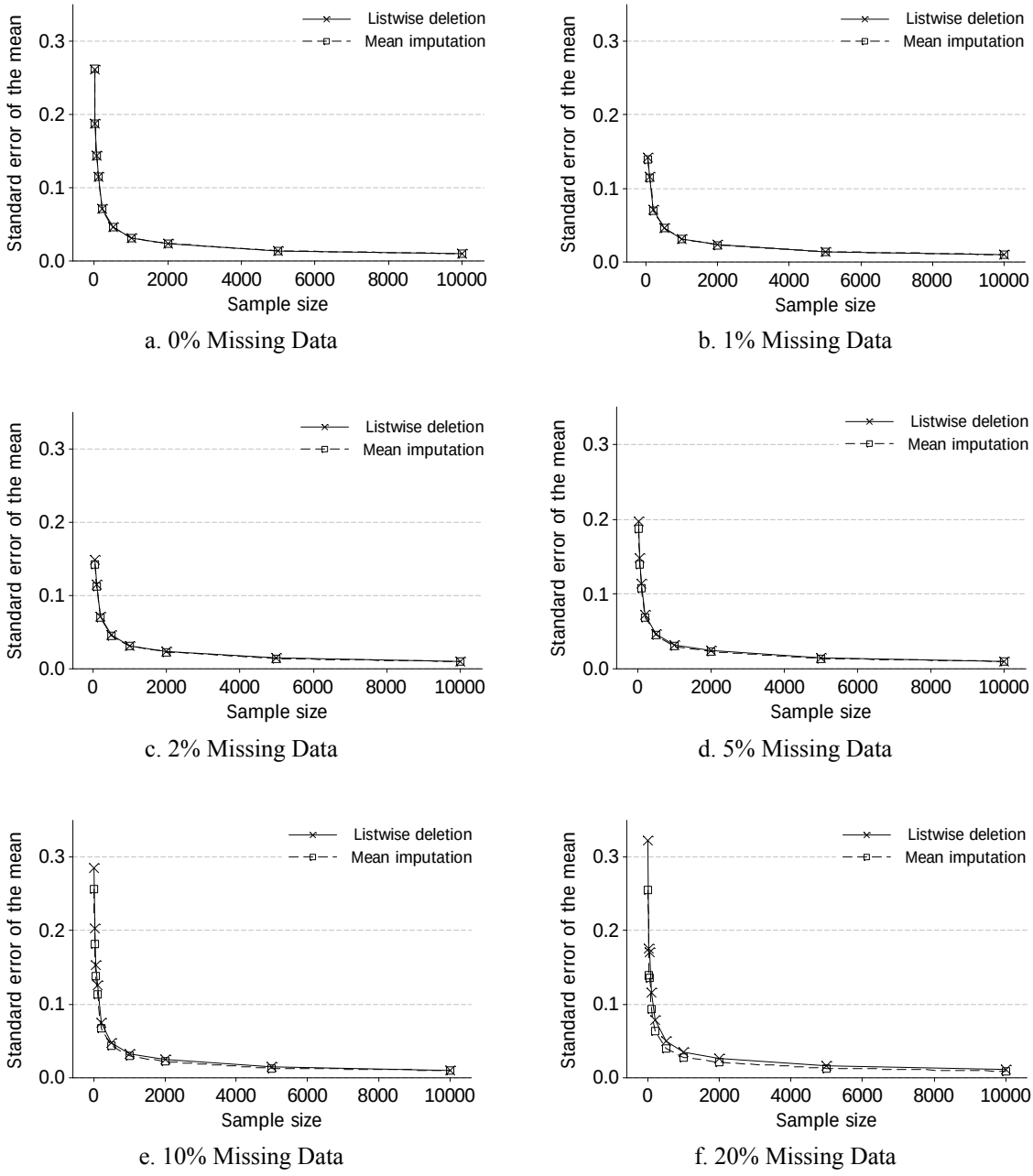
		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
SE	10	0.262	-	-	-	0.286	0.323
	20	0.188	-	-	0.197	0.203	0.176
	50	0.144	0.142	0.149	0.149	0.153	0.170
	100	0.115	0.115	0.116	0.113	0.126	0.116
	200	0.071	0.071	0.072	0.072	0.074	0.078
	500	0.046	0.046	0.046	0.047	0.048	0.050
	1000	0.031	0.031	0.031	0.032	0.033	0.035
	2000	0.024	0.024	0.024	0.024	0.025	0.026
	5000	0.014	0.014	0.014	0.015	0.015	0.016
	10000	0.010	0.010	0.010	0.010	0.011	0.011
	Mean	0.090	0.057	0.058	0.073	0.097	0.100
Relative Error of *SE*	10	-	-	-	-	0.092	0.234
	20	-	-	-	0.051	0.081	-0.065
	50	-	-0.010	0.038	0.035	0.065	0.187
	100	-	0.007	0.010	-0.011	0.098	0.014
	200	-	0.006	0.014	0.021	0.050	0.106
	500	-	0.002	0.000	0.019	0.037	0.087
	1000	-	0.006	0.010	0.026	0.055	0.126
	2000	-	0.004	0.004	0.026	0.051	0.115
	5000	-	0.007	0.014	0.028	0.049	0.127
	10000	-	0.010	0.010	0.030	0.050	0.110
	Mean	-	0.004	0.013	0.025	0.063	0.104

Note. Mean relative error of *SE* = 0.045; *RMSE* = 0.011 .

Table 12. Standard Error of the Mean and its Relative Error under One Sample *t* Test with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
SE	10	0.262	-	-	-	0.256	0.255
	20	0.188	-	-	0.187	0.182	0.140
	50	0.144	0.139	0.143	0.140	0.138	0.136
	100	0.115	0.114	0.113	0.108	0.113	0.093
	200	0.071	0.070	0.070	0.069	0.067	0.063
	500	0.046	0.046	0.045	0.045	0.043	0.040
	1000	0.031	0.031	0.031	0.030	0.029	0.028
	2000	0.024	0.023	0.023	0.023	0.022	0.021
	5000	0.014	0.014	0.014	0.014	0.013	0.013
	10000	0.010	0.010	0.010	0.010	0.010	0.009
	Mean	0.091	0.056	0.056	0.070	0.087	0.080
Relative Error of *SE*	10	-	-	-	-	-0.023	-0.027
	20	-	-	-	-0.005	-0.032	-0.255
	50	-	-0.035	-0.007	-0.028	-0.042	-0.056
	100	-	-0.009	-0.017	-0.061	-0.017	-0.191
	200	-	-0.014	-0.014	-0.028	-0.056	-0.113
	500	-	0.000	-0.022	-0.022	-0.065	-0.130
	1000	-	0.000	0.000	-0.032	-0.065	-0.097
	2000	-	-0.042	-0.042	-0.042	-0.083	-0.125
	5000	-	0.000	0.000	0.000	-0.071	-0.071
	10000	-	0.000	0.000	0.000	0.000	-0.100
	Mean	-	-0.012	-0.013	-0.024	-0.045	-0.117

Note. Mean relative error of *SE* = -0.045; *RMSE* = 0.009 .



a. 0% Missing Data

b. 1% Missing Data

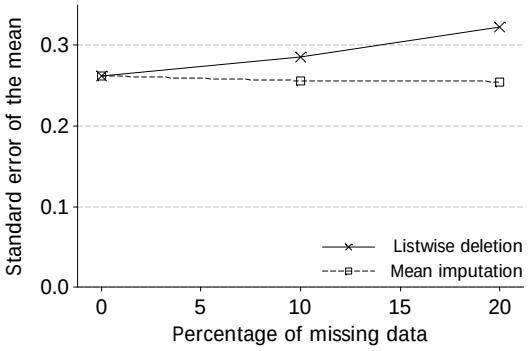
c. 2% Missing Data

d. 5% Missing Data

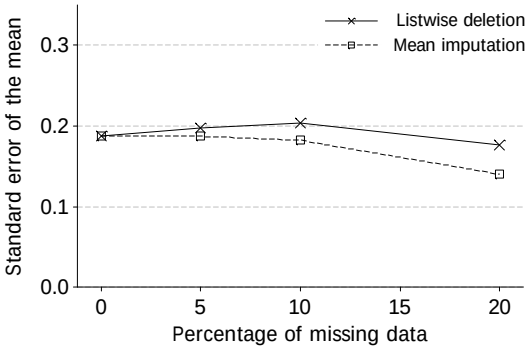
e. 10% Missing Data

f. 20% Missing Data

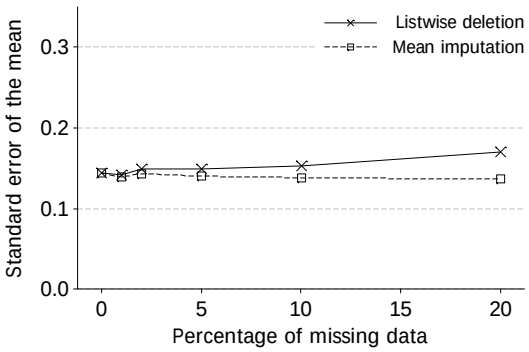
Figure 17. Standard error of sample mean plotted as a function of sample size for missing sample data ranging between 0% and 20% using one sample *t* test with listwise deletion and mean imputation.



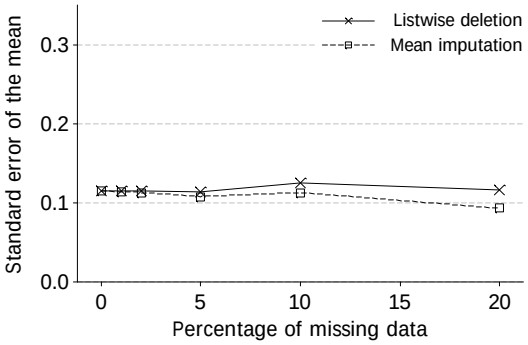
a. $n = 10$



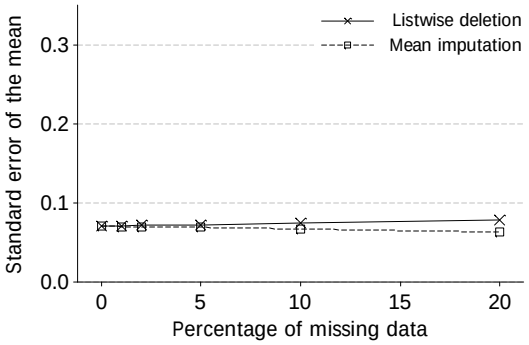
b. $n = 20$



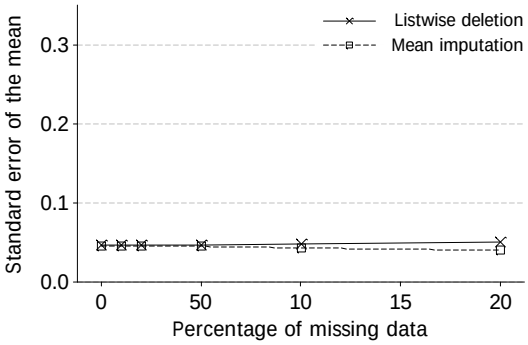
c. $n = 50$



d. $n = 100$



e. $n = 200$



f. $n = 500$

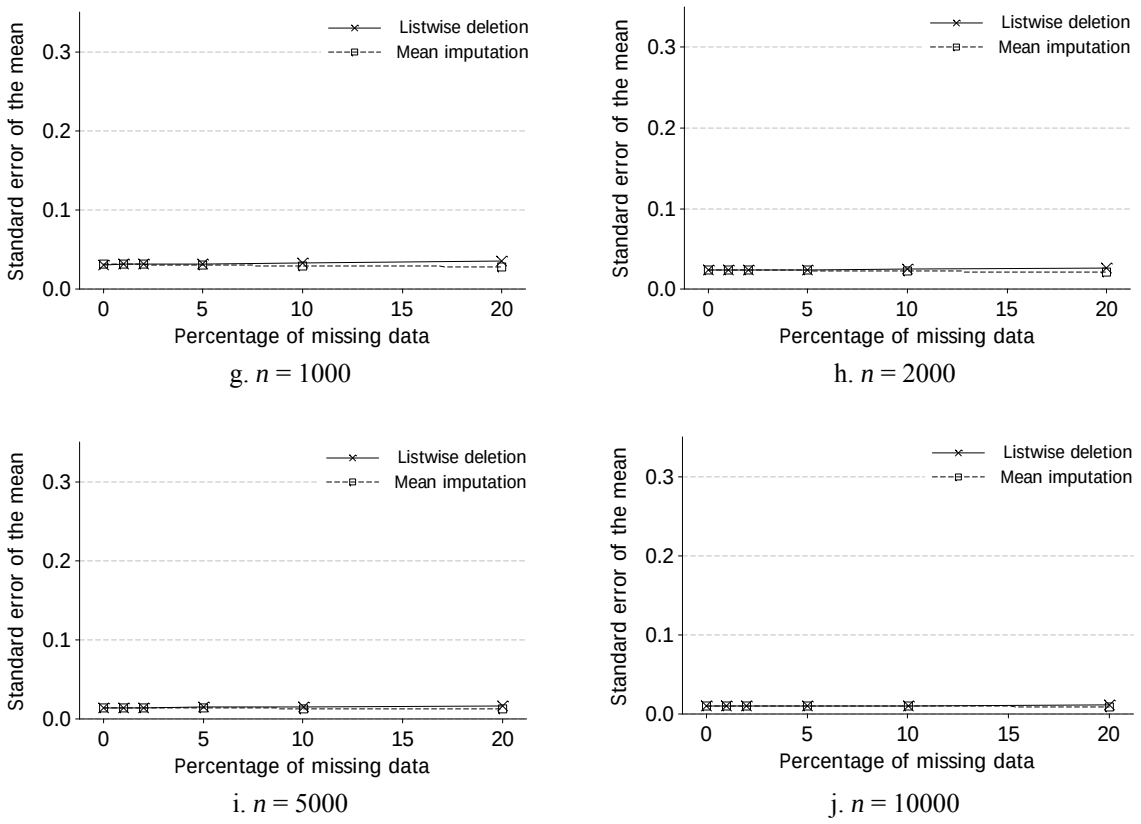
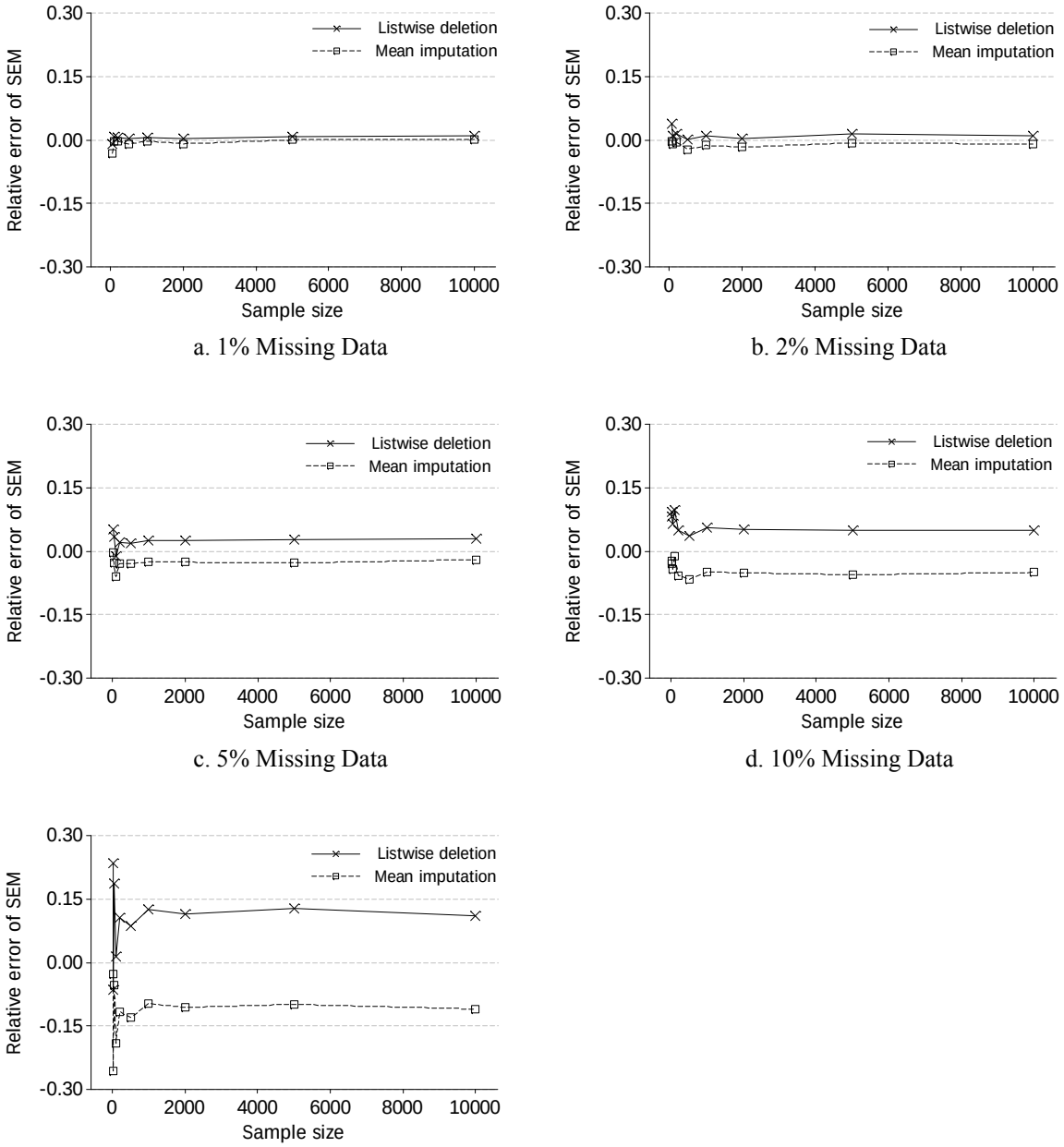


Figure 18. Standard error of sample mean plotted as a function of percentage of missing data for sample size ranging between 10 and 10,000 under one sample *t* test with listwise deletion and mean imputation.

show that listwise deletion works relatively better than mean imputation when the sample size is small because standard errors under listwise deletion are larger than those under mean imputation (even though they are still smaller than those in complete data). However, this difference in standard errors becomes negligible as sample size becomes large. For example, we see that in panel *j* of Figure 18, the two standard error curves are almost identical.

The difference in performance of listwise deletion and mean imputation is easier to see when relative error of *SE* is plotted against sample size in Figure 19 and against proportion of missing data in Figure 20. Figure 19 shows that although relative error of *SE* remains constant for large sample sizes, there is a persistent gap between relative error of *SE* based on listwise deletion and mean imputation, with the gap increasing as percentage of missing data increases. In other words although both listwise deletion and mean imputation underestimate SE, the performance of mean imputation deteriorates more rapidly as compared to listwise deletion as proportion of missing data increases. This difference in performance between the two methods can be seen with more clarity in Figure 20 where relative error of SE is plotted against proportion of missing data for various sample sizes. Figure 20 makes it very clear that, regardless of sample size, the gap in performance of the two methods, listwise deletion and mean imputation, increases rapidly as proportion of missing data increases.

The observed values of *t* statistic and its relative error obtained by dividing sample means in Table 10 by the standard errors under listwise deletion in Table 11 are presented in Table 13 with the corresponding *p* values for the observed *t* statistic presented in Table 14. The observed values of *t* statistic and its relative error obtained by dividing sample means in Table 10 by the standard errors under mean imputation in Table 12 are presented in Table 15 with the corresponding *p* values for the observed *t* statistic presented in Table 16. The *t* value is a better criterion for evaluating the comparative performance of missing data handling methods because it is not the sample mean nor its standard error but a combination of both in the form of *t* statistic that



a. 1% Missing Data

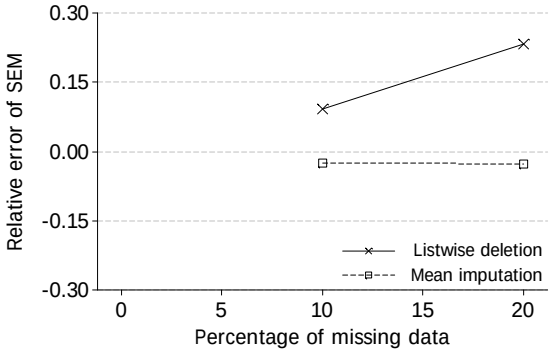
b. 2% Missing Data

c. 5% Missing Data

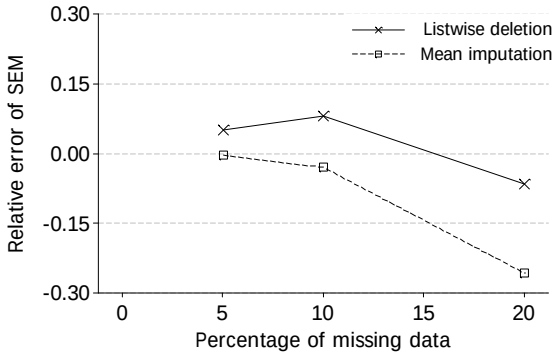
d. 10% Missing Data

e. 20% Missing Data

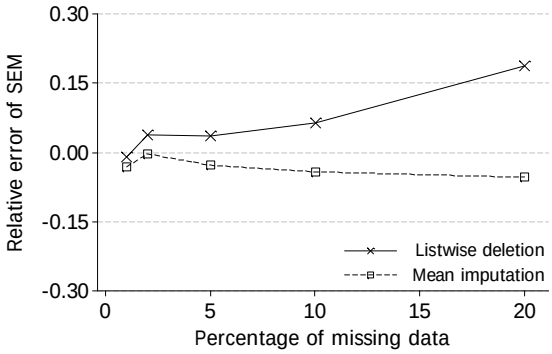
Figure 19. Relative error of standard error of sample mean (SEM) plotted as a function of percentage of missing data for sample size ranging between 10 and 10,000 under one sample *t* test with listwise deletion and mean imputation.



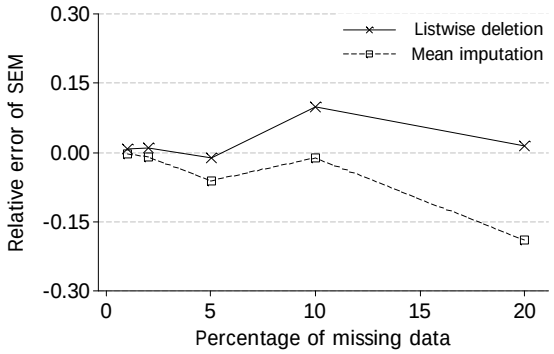
a. $n = 10$



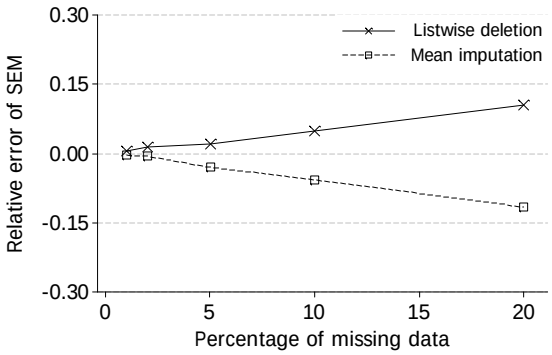
b. $n = 20$



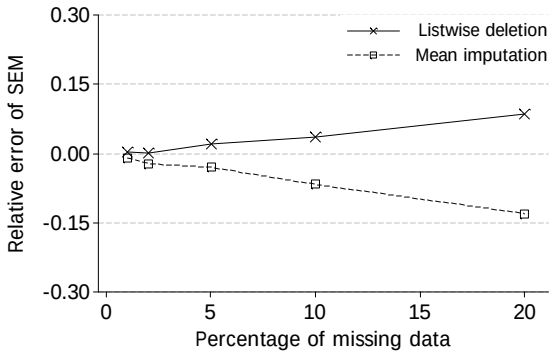
c. $n = 50$



d. $n = 100$



e. $n = 200$



f. $n = 500$

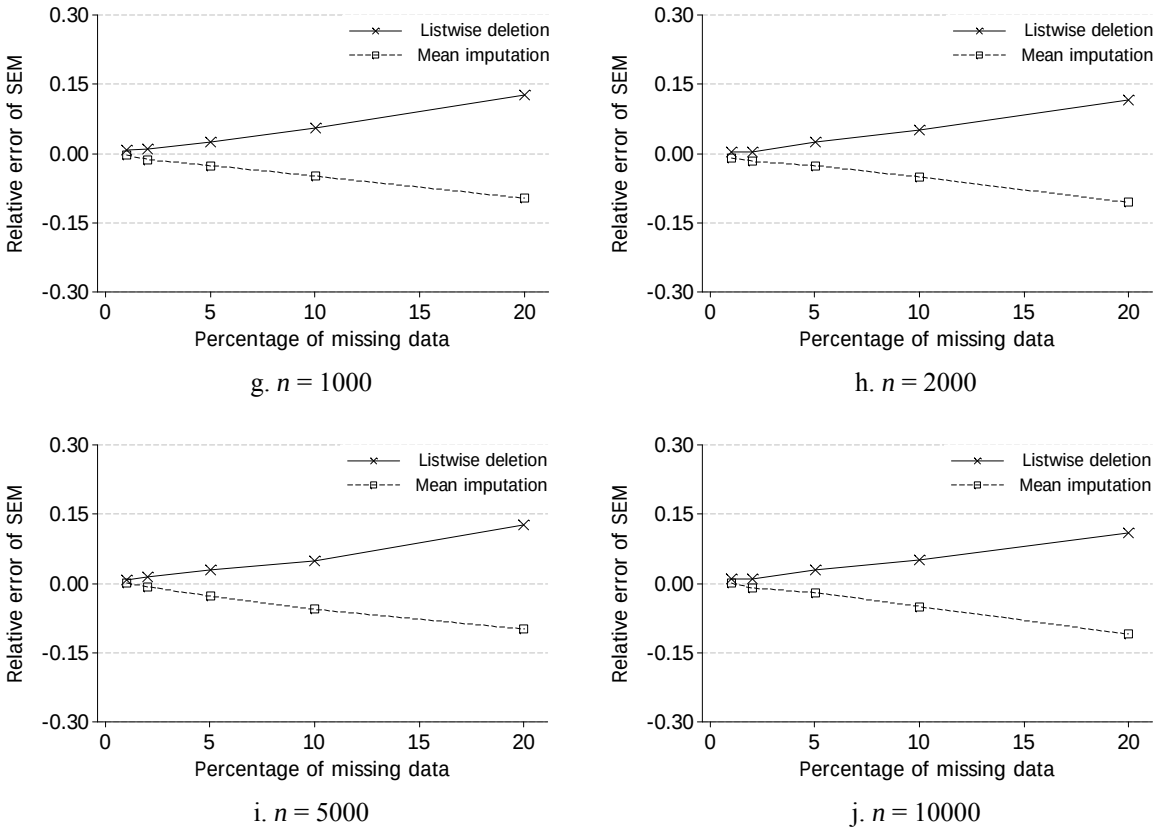


Figure 20. Relative error of standard error of sample mean (SEM) plotted as a function of percentage of missing data for sample size ranging between 10 and 10,000 under one sample *t* test with listwise deletion and mean imputation.

determines whether we make the correct or incorrect decision in rejecting or not rejecting the null hypothesis. For this reason, from this point onwards this study evaluates the overall effect of a method of analysis by looking at the model significance test statistic rather than individual parameter estimates as the number of such estimates increases very fast with the increase in complexity of the model being tested. For example, the linear multiple regression model with four predictors discussed elsewhere in this section

Table 13. Observed Test Statistic and its Relative Error under One Sample *t* Test with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.335	-	-	-	-0.504	-0.376
	20	0.321	-	-	0.388	0.019	0.538
	50	-0.236	0.009	-0.313	-0.281	-0.458	-0.009
	100	-0.631	-0.543	-0.419	-0.658	-0.722	-0.957
	200	0.673	0.558	0.507	0.814	-0.046	0.45
	500	0.419	0.479	0.035	0.699	0.928	-0.386
	1000	0.564	0.615	0.531	0.559	0.274	0.873
	2000	0.292	0.346	0.284	0.058	0.471	-0.336
	5000	0.851	0.975	1.053	0.701	1.383	-0.036
	10000	0.978	1.043	0.903	0.928	1.055	0.968
	Mean	0.290	0.435	0.323	0.356	0.240	0.073
Relative Error of *t*	10	-	-	-	-	0.505	0.122
	20	-	-	-	0.209	-0.941	0.676
	50	-	-1.038	0.326	0.191	0.941	-0.962
	100	-	-0.140	-0.336	0.043	0.144	0.517
	200	-	-0.171	-0.247	0.210	-1.068	-0.331
	500	-	0.143	-0.917	0.668	1.215	-1.921
	1000	-	0.090	-0.059	-0.009	-0.514	0.548
	2000	-	0.185	-0.027	-0.801	0.613	-2.151
	5000	-	0.146	0.237	-0.176	0.625	-1.042
	10000	-	0.067	-0.077	-0.051	0.079	-0.010
	Mean	-	-0.090	-0.138	0.032	0.160	-0.455

Note. Mean relative error of *t* = -0.100; *RMSE* = 0.302 .

Table 14. Observed Probability of Type I Error under One Sample *t* Test with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.745	-	-	-	.628	.718
	20	.752	-	-	.702	.985	.599
	50	.815	.993	.756	.780	.649	.993
	100	.530	.588	.676	.512	.472	.342
	200	.502	.578	.613	.417	.963	.654
	500	.676	.632	.972	.485	.354	.700
	1000	.573	.539	.596	.576	.784	.383
	2000	.771	.730	.777	.953	.638	.737
	5000	.395	.330	.293	.484	.167	.971
	10000	.328	.297	.367	.354	.292	.333
	Mean	.609	.586	.631	.585	.593	.643

produces one intercept and four partial slope parameter estimates along with the standard errors of these five estimates for each of the ten sample sizes at each of the five proportions of missing data considered in his study. This example clearly illustrates the complexity that will be introduced in our analysis by focusing on parameter estimates and their standard errors rather than the global model test statistics. Since the four methods of analysis considered in this study use different model specifications and have different parameters, in addition to the increase in complexity, a focus on parameters will make it very difficult to compare the performance of various missing data handling methods across those methods of analysis.

Table 15. Observed Test Statistic and its Relative Error under One Sample *t* Test with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.335	-	-	-	-0.563	-0.476
	20	0.321	-	-	0.409	0.021	0.677
	50	-0.236	0.009	-0.326	-0.299	-0.510	-0.011
	100	-0.631	-0.549	-0.428	-0.693	-0.803	-1.197
	200	0.673	0.563	0.518	0.857	-0.052	0.562
	500	0.419	0.484	0.036	0.736	1.031	-0.483
	1000	0.564	0.621	0.542	0.589	0.305	1.091
	2000	0.292	0.349	0.290	0.062	0.523	-0.420
	5000	0.851	0.985	1.074	0.738	1.537	-0.046
	10000	0.978	1.054	0.921	0.976	1.172	1.210
	Mean	0.290	0.440	0.328	0.375	0.266	0.091
Relative Error of *t*	10	-	-	-	-	0.681	0.421
	20	-	-	-	0.274	-0.935	1.109
	50	-	-1.038	0.381	0.267	1.161	-0.953
	100	-	-0.130	-0.322	0.098	0.273	0.897
	200	-	-0.163	-0.230	0.273	-1.077	-0.165
	500	-	0.155	-0.914	0.757	1.461	-2.153
	1000	-	0.101	-0.039	0.044	-0.459	0.934
	2000	-	0.195	-0.007	-0.788	0.791	-2.438
	5000	-	0.157	0.262	-0.133	0.806	-1.054
	10000	-	0.078	-0.058	-0.002	0.198	0.237
	Mean	-	-0.081	-0.116	0.088	0.290	-0.316

Note. Mean relative error of t = -0.023; *RMSE* = 0.345 .

Table 16. Observed Probability of Type I Error under One Sample *t* Test with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.745	-	-	-	.587	.645
	20	.752	-	-	.687	.983	.507
	50	.815	.993	.746	.766	.613	.992
	100	.530	.585	.670	.490	.424	.234
	200	.502	.574	.605	.393	.959	.575
	500	.676	.629	.971	.462	.303	.629
	1000	.573	.535	.588	.556	.761	.276
	2000	.771	.727	.772	.951	.601	.675
	5000	.395	.325	.283	.461	.124	.964
	10000	.328	.292	.357	.329	.241	.226
	Mean	.609	.583	.624	.566	.560	.572

It is worthwhile to point here that another statistic that is widely used to compare performance of competing models under a variety of circumstances is the effect size. For simulated data, computation of effect size estimates does not make much sense because the relationships between variables are completely hypothetical, are not supported by any underlying theoretical framework, and the purpose of deploying the various methods of analysis is not to explain variation in the dependent variable. The situation is however different for the two empirical datasets used for illustration where the effect sizes are indeed meaningful.

In order to compare the overall performance of listwise deletion with mean imputation, the observed *t* values from Tables 13 and 15 are plotted against each other in Figure 21 while their corresponding *p* values from Tables 14 and 16 are plotted against each other in Figure 22. Figures 21 and 22 show that the overall performance of both missing data handling methods is very similar with the correlation between their observed statistics and corresponding p values being in excess of .99 . The overall conclusion here is that both listwise deletion and mean imputation produce very similar results in terms of their effect on the test of hypothesis in a single sample *t* test (although this may not be true for other methods of analysis). As a final test, the mean absolute *t* values are plotted as a function of proportion of missing data in Figure 23.

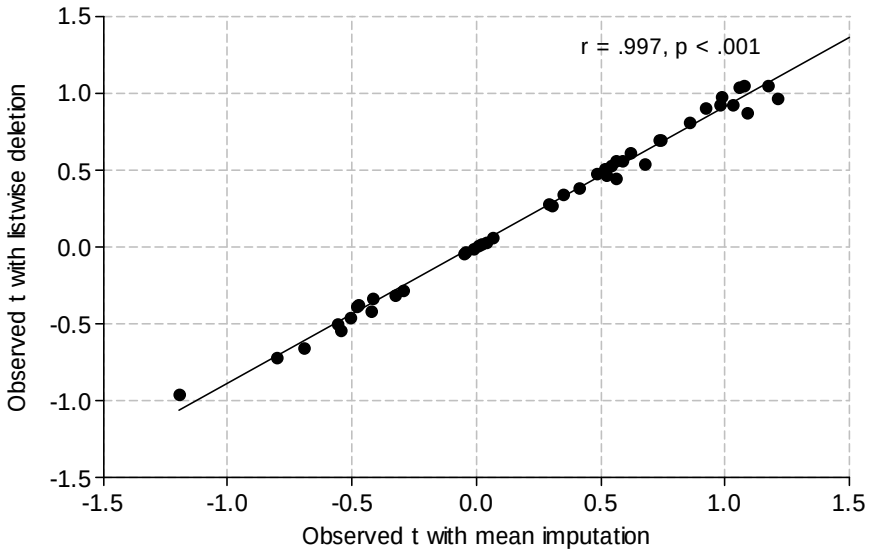


Figure 21. Linear regression line between *t* values obtained from listwise deletion and those obtained from mean imputation when data is missing. Regression equation is $\hat{Y} = .01 + .90X$.

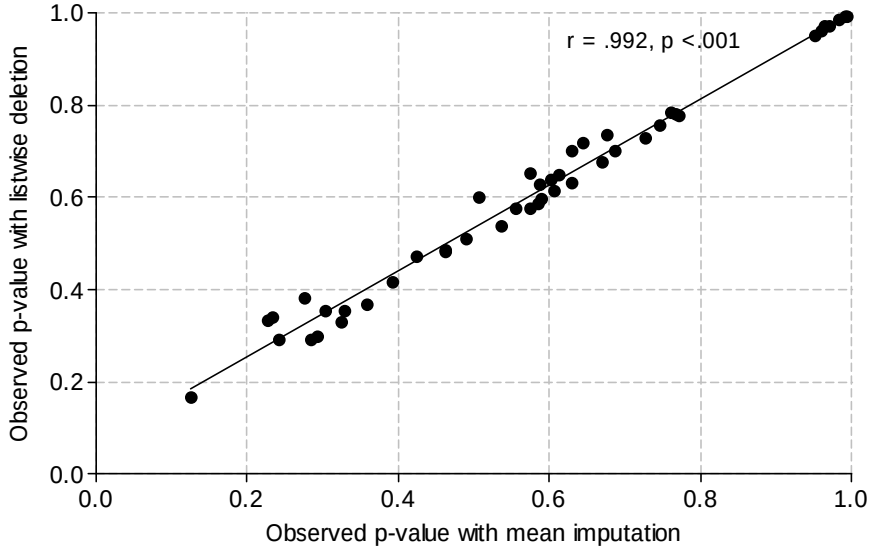


Figure 22. Linear regression line between *p*-values obtained from listwise deletion and those obtained from mean imputation when data is missing. Regression equation is $\hat{Y} = .07 + .93X$.

It can be seen from Figure 23 that although there is a difference between the two methods in their effect on *t*, and even though this difference increases as the proportion of missing data increases, the mean *t* value remains within its 95% confidence interval. In other words, both methods end up generating the same decision for the test of hypothesis, $H_o : \mu_Y = 0$. The observed values of *t* statistic and its relative error are presented in Tables 17, 19, and 21 with the corresponding *p* value for the *t* statistic presented in Tables 18, 20, and 22 under regression imputation, EM imputation, and multiple imputation respectively. Power analysis for 10 complete samples and 50 samples containing missing data for one sample *t* test under the five missing data handling methods is provided in

Table 23. Observed *t* statistic along with its relative error, corresponding *p* values, and power analysis are provided in Tables 24 through 34 for independent samples *t* test, in Tables 35 through 45 for two-way ANOVA, and in Tables 46 through 56 for linear multiple regression.

Rather than indulge in a lengthy description of the large volume of statistics presented in tables identified in the preceding paragraph, it is relatively more efficient to discuss a summary of their findings. Readers who are interested in very specific information regarding a method of analysis, missing data handling method, sample size, or proportion of missing data can still find that information in the raw tables. Information presented in Tables 24 through 33, 35 through 44, and 46 through 55 is summarized in Figures 24 through 28.

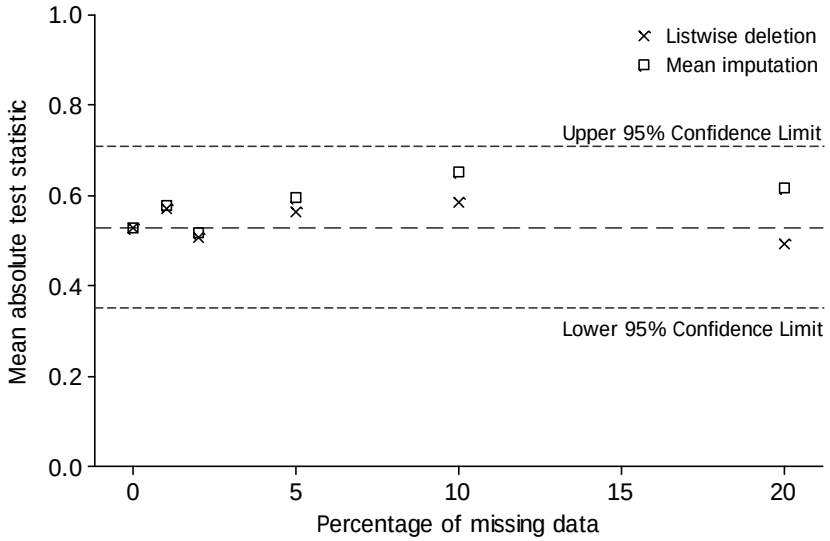


Figure 23. Absolute test statistic values averaged over various sample sizes plotted as a function of percentage of missing data under one sample *t* test using listwise deletion and mean imputation. The horizontal reference lines are for mean absolute *t* and the corresponding 95% confidence limits when 0% of the data is missing.

Table 17. Observed Test Statistic and its Relative Error under One Sample *t* Test with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.335	-	-	-	-0.952	-0.017
	20	0.321	-	-	0.161	0.085	0.445
	50	-0.236	-0.249	-0.204	-0.101	-0.686	-0.116
	100	-0.631	-0.753	-0.674	-0.736	-0.927	-0.744
	200	0.673	0.711	0.500	0.757	0.780	-0.055
	500	0.419	0.528	0.396	0.420	0.714	0.111
	1000	0.564	0.585	0.543	0.566	0.801	0.970
	2000	0.292	0.277	0.189	0.026	0.921	-0.442
	5000	0.851	0.983	0.981	1.198	1.186	0.583
	10000	0.978	1.020	0.921	1.198	1.333	1.647
	Mean	0.290	0.388	0.332	0.388	0.326	0.238
Relative Error of *t*	10	-	-	-	-	1.842	-0.949
	20	-	-	-	-0.498	-0.735	0.386
	50	-	0.055	-0.136	-0.572	1.907	-0.508
	100	-	0.193	0.068	0.166	0.469	0.179
	200	-	0.056	-0.257	0.125	0.159	-1.082
	500	-	0.260	-0.055	0.002	0.704	-0.735
	1000	-	0.037	-0.037	0.004	0.420	0.720
	2000	-	-0.051	-0.353	-0.911	2.154	-2.514
	5000	-	0.155	0.153	0.408	0.394	-0.315
	10000	-	0.043	-0.058	0.225	0.363	0.684
	Mean	-	0.094	-0.084	-0.117	0.768	-0.413

Note. Mean relative error of $t = 0.057$; $RMSE = 0.295$.

Table 18. Observed Probability of Type I Error under One Sample *t* Test with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.745	-	-	-	.366	.987
	20	.752	-	-	.874	.933	.661
	50	.815	.804	.839	.900	.496	.908
	100	.530	.453	.502	.463	.356	.458
	200	.502	.478	.618	.450	.436	.956
	500	.676	.598	.692	.675	.475	.911
	1000	.573	.559	.587	.572	.423	.332
	2000	.771	.782	.850	.979	.357	.659
	5000	.395	.326	.327	.231	.236	.560
	10000	.328	.308	.357	.231	.183	.100
	Mean	.609	.539	.597	.597	.426	.653

Table 19. Observed Test Statistic and its Relative Error under One Sample *t* Test with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.335	-	-	-	-0.750	-0.411
	20	0.321	-	-	0.374	-0.004	0.708
	50	-0.236	-0.227	-0.107	-0.157	-0.388	-0.051
	100	-0.631	-0.692	-0.648	-0.839	-0.784	-0.745
	200	0.673	0.683	0.635	0.759	0.266	0.379
	500	0.419	0.509	0.341	0.436	0.673	0.056
	1000	0.564	0.587	0.516	0.475	0.603	1.125
	2000	0.292	0.333	0.234	0.345	0.798	-0.136
	5000	0.851	0.952	0.937	0.814	1.101	0.816
	10000	0.978	1.011	0.870	1.151	1.365	1.354
	Mean	0.290	0.395	0.347	0.373	0.288	0.310
Relative Error of *t*	10	-	-	-	-	1.239	0.227
	20	-	-	-	0.165	-1.012	1.206
	50	-	-0.038	-0.547	-0.335	0.644	-0.784
	100	-	0.097	0.027	0.330	0.242	0.181
	200	-	0.015	-0.056	0.128	-0.605	-0.437
	500	-	0.215	-0.186	0.041	0.606	-0.866
	1000	-	0.041	-0.085	-0.158	0.069	0.995
	2000	-	0.140	-0.199	0.182	1.733	-1.466
	5000	-	0.119	0.101	-0.043	0.294	-0.041
	10000	-	0.034	-0.110	0.177	0.396	0.384
	Mean	-	0.078	-0.132	0.054	0.361	-0.060

Note. Mean relative error of $t = 0.068$; $RMSE = 0.224$.

Table 20. Observed Probability of Type I Error under One Sample *t* Test with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.745	-	-	-	.472	.691
	20	.752	-	-	.713	.997	.488
	50	.815	.821	.916	.876	.699	.959
	100	.530	.491	.518	.404	.435	.458
	200	.502	.496	.526	.449	.791	.705
	500	.676	.611	.733	.663	.501	.956
	1000	.573	.557	.606	.635	.546	.261
	2000	.771	.739	.815	.730	.425	.892
	5000	.395	.341	.349	.416	.271	.414
	10000	.328	.312	.384	.250	.172	.176
	Mean	.609	.546	.606	.571	.531	.600

Table 21. Observed Test Statistic and its Relative Error under One Sample *t* Test with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.335	-	-	-	-0.243	-0.037
	20	0.321	-	-	0.425	0.027	0.732
	50	-0.236	-0.196	-0.277	-0.086	-0.223	0.139
	100	-0.631	-0.728	-0.707	-0.794	-0.608	-0.705
	200	0.673	0.695	0.639	0.644	0.462	0.436
	500	0.419	0.454	0.329	0.472	0.658	-0.002
	1000	0.564	0.600	0.450	0.389	0.472	0.803
	2000	0.292	0.323	0.251	0.303	0.616	0.182
	5000	0.851	0.938	0.849	0.978	1.099	0.880
	10000	0.978	0.945	0.872	1.061	1.139	1.281
	Mean	0.290	0.379	0.301	0.377	0.340	0.371
Relative Error of *t*	10	-	-	-	-	-0.275	-0.890
	20	-	-	-	0.324	-0.916	1.280
	50	-	-0.169	0.174	-0.636	-0.055	-1.589
	100	-	0.154	0.120	0.258	-0.036	0.117
	200	-	0.033	-0.051	-0.043	-0.314	-0.352
	500	-	0.084	-0.215	0.126	0.570	-1.005
	1000	-	0.064	-0.202	-0.310	-0.163	0.424
	2000	-	0.106	-0.140	0.038	1.110	-0.377
	5000	-	0.102	-0.002	0.149	0.291	0.034
	10000	-	-0.034	-0.108	0.085	0.165	0.310
	Mean	-	0.042	-0.053	-0.001	0.038	-0.205

Note. Mean relative error of *t* = -0.039; *RMSE* = 0.175 .

Table 22. Observed Probability of Type I Error under One Sample *t* Test with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.745	-	-	-	.813	.971
	20	.752	-	-	.676	.979	.473
	50	.814	.845	.783	.932	.824	.890
	100	.529	.468	.481	.429	.545	.482
	200	.502	.488	.524	.520	.645	.663
	500	.675	.650	.742	.637	.511	.998
	1000	.573	.549	.653	.697	.637	.422
	2000	.770	.747	.802	.762	.538	.856
	5000	.395	.348	.396	.328	.272	.379
	10000	.328	.345	.383	.289	.255	.200
	Mean	.608	.555	.596	.586	.602	.633

Table 23. Power of the Test to Detect Medium Effect Size under One Sample *t* Test for Various Missing Data Handling Methods.

		Percentage of Missing Data					
Method	*n*	0%	1%	2%	5%	10%	20%
Listwise Deletion	10	.293	-	-	-	.263	.232
	20	.565	-	-	.541	.516	.465
	50	.933	.929	.924	.919	.907	.869
	100	.999	.998	.998	.998	.997	.993
	200	1.000	1.000	1.000	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000
Mean Imputation Regression Imputation EM Imputation Multiple imputation	10	.293	.293	.293	.293	.293	.293
	20	.565	.565	.565	.565	.565	.565
	50	.933	.933	.933	.933	.933	.933
	100	.999	.999	.999	.999	.999	.999
	200	1.000	1.000	1.000	1.000	1.000	1.000
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000

Note. EM = Expectation Maximization. Medium effect size defined as Cohen's $d = .5$. Power calculations assume a two-tailed test with $\alpha = .05$

Table 24. Observed Test Statistic and its Relative Error under Independent Samples *t* Test with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.288	-	-	-	-0.695	-0.703
	20	-0.127	-	-	-0.216	-0.403	0.269
	50	0.902	0.682	0.814	0.636	0.696	0.919
	100	0.627	0.714	0.597	0.653	0.586	0.389
	200	-0.107	0.010	-0.215	0.247	-0.268	-0.067
	500	-0.257	-0.029	-0.289	-0.808	-0.699	-0.628
	1000	0.706	0.670	0.606	0.399	0.629	1.152
	2000	-0.311	-0.208	-0.247	0.040	-0.321	-0.508
	5000	-0.117	-1.144	-1.184	-0.839	-1.380	-1.445
	10000	0.002	0.090	-0.060	-0.097	-0.168	0.913
	Mean	0.103	0.098	0.003	0.002	-0.202	0.029
Relative Error of *t*	10	-	-	-	-	1.413	1.441
	20	-	-	-	0.701	2.173	-3.118
	50	-	-0.244	-0.098	-0.295	-0.228	0.019
	100	-	0.139	-0.048	0.041	-0.065	-0.380
	200	-	-1.093	1.009	-3.308	1.505	-0.374
	500	-	-0.887	0.125	2.144	1.720	1.444
	1000	-	-0.051	-0.142	-0.435	-0.109	0.632
	2000	-	-0.331	-0.206	-1.129	0.032	0.633
	5000	-	8.778	9.120	6.171	10.795	11.350
	10000	-	44.000	-31.000	-49.500	-85.000	455.500
	Mean	-	6.289	-2.655	-5.068	-6.776	46.715

Note. Mean relative error of *t* = 8.508; *RMSE* = 0.450 .

Table 25. Observed Probability of Type I Error under Independent Samples *t* Test with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.781	-	-	-	.510	.510
	20	.901	-	-	.832	.692	.794
	50	.371	.498	.420	.528	.490	.364
	100	.532	.477	.552	.516	.559	.698
	200	.915	.992	.830	.805	.789	.947
	500	.797	.977	.773	.419	.485	.531
	1000	.481	.503	.545	.690	.529	.250
	2000	.756	.836	.805	.968	.749	.612
	5000	.264	.253	.237	.402	.168	.149
	10000	.998	.929	.952	.923	.866	.361
	Mean	.680	.683	.639	.676	.584	.522

Table 26. Observed Test Statistic and its Relative Error under Independent Samples *t* Test with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.288	-	-	-	-0.618	-0.683
	20	-0.127	-	-	-0.220	-0.398	0.276
	50	0.902	0.684	0.811	0.638	0.699	0.920
	100	0.627	0.714	0.597	0.647	0.586	0.388
	200	-0.107	0.010	-0.215	0.247	-0.268	-0.067
	500	-0.257	-0.029	-0.289	-0.808	-0.699	-0.628
	1000	0.706	0.670	0.606	0.399	0.630	1.152
	2000	-0.311	-0.208	-0.247	0.040	-0.321	-0.508
	5000	-1.117	-1.144	-1.184	-0.839	-1.380	-1.445
	10000	0.002	0.090	-0.060	-0.097	-0.168	0.913
	Mean	0.003	0.098	0.002	0.001	-0.194	0.032
Relative Error of *t*	10	-	-	-	-	1.146	1.372
	20	-	-	-	0.732	2.134	-3.173
	50	-	-0.242	-0.101	-0.293	-0.225	0.020
	100	-	0.139	-0.048	0.032	-0.065	-0.381
	200	-	-1.093	1.009	-3.308	1.505	-0.374
	500	-	-0.887	0.125	2.144	1.720	1.444
	1000	-	-0.051	-0.142	-0.435	-0.108	0.632
	2000	-	-0.331	-0.206	-1.129	0.032	0.633
	5000	-	0.024	0.060	-0.249	0.235	0.294
	10000	-	44.000	-31.000	-49.500	-85.000	455.500
	Mean	-	5.195	-3.788	-5.778	-7.863	45.597

Note. Mean relative error of t = 7.480; *RMSE* = 0.268 .

Table 27. Observed Probability of Type I Error under Independent Samples *t* Test with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.781	-	-	-	.557	.517
	20	.901	-	-	.829	.696	.787
	50	.371	.497	.421	.526	.488	.362
	100	.532	.477	.552	.519	.559	.699
	200	.915	.992	.830	.805	.789	.946
	500	.797	.977	.773	.420	.485	.530
	1000	.481	.503	.545	.690	.529	.249
	2000	.756	.836	.805	.968	.749	.612
	5000	.264	.253	.237	.402	.168	.149
	10000	.998	.929	.952	.923	.866	.361
	Mean	.680	.683	.639	.676	.589	.521

Table 28. Observed Test Statistic and its Relative Error under Independent Samples *t* Test with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.288	-	-	-	-1.148	-1.067
	20	-0.127	-	-	0.038	-0.325	0.419
	50	0.902	0.913	0.924	0.496	0.501	0.846
	100	0.627	0.479	0.412	0.515	0.904	0.531
	200	-0.107	-0.148	-0.069	-0.232	-0.127	-0.694
	500	-0.257	-0.230	-0.334	-0.658	-0.568	-1.102
	1000	0.706	0.729	0.581	0.509	0.448	0.205
	2000	-0.311	-0.223	-0.375	-0.296	-0.611	-0.305
	5000	-1.117	-1.298	-1.353	-1.189	-1.354	-1.275
	10000	0.002	0.100	-0.054	0.093	-0.124	0.058
	Mean	0.003	0.040	-0.034	-0.080	-0.240	-0.238
Relative Error of *t*	10	-	-	-	-	2.986	2.705
	20	-	-	-	-1.299	1.559	-4.299
	50	-	0.012	0.024	-0.450	-0.445	-0.062
	100	-	-0.236	-0.343	-0.179	0.442	-0.153
	200	-	0.383	-0.355	1.168	0.187	5.486
	500	-	-0.105	0.300	1.560	1.210	3.288
	1000	-	0.033	-0.177	-0.279	-0.365	-0.710
	2000	-	-0.283	0.206	-0.048	0.965	-0.019
	5000	-	0.162	0.211	0.064	0.212	0.141
	10000	-	49.000	-28.000	45.500	-63.000	28.000
	Mean	-	6.121	-3.517	5.115	-5.625	3.438

Note. Mean relative error of t = 1.000; *RMSE* = 0.307 .

Table 29. Observed Probability of Type I Error under Independent Samples *t* Test with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.781	-	-	-	.284	.323
	20	.901	-	-	.970	.750	.682
	50	.371	.366	.360	.622	.619	.402
	100	.532	.633	.681	.608	.368	.597
	200	.915	.883	.945	.817	.899	.488
	500	.797	.818	.738	.511	.570	.271
	1000	.481	.466	.561	.611	.654	.838
	2000	.756	.824	.708	.767	.541	.760
	5000	.264	.194	.176	.235	.176	.202
	10000	.998	.920	.957	.926	.901	.954
	Mean	.680	.638	.641	.674	.576	.552

Table 30. Observed Test Statistic and its Relative Error under Independent Samples *t* Test with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.288	-	-	-	-0.895	-1.211
	20	-0.127	-	-	-0.182	-0.410	0.180
	50	0.902	0.895	1.029	0.731	0.780	0.959
	100	0.627	0.557	0.514	0.449	0.727	0.270
	200	-0.107	-0.117	-0.086	-0.075	-0.364	-0.307
	500	-0.257	-0.195	-0.258	-0.527	-0.492	-0.652
	1000	0.706	0.761	0.592	0.505	0.681	0.474
	2000	-0.311	-0.231	-0.278	-0.278	-0.417	-0.390
	5000	-1.117	-1.156	-1.265	-0.992	-1.214	-1.481
	10000	0.002	0.079	-0.111	0.144	-0.180	-0.247
	Mean	0.003	0.074	0.017	-0.025	-0.178	-0.241
Relative Error of *t*	10	-	-	-	-	2.108	3.205
	20	-	-	-	0.433	2.228	-2.417
	50	-	-0.008	0.141	-0.190	-0.135	0.063
	100	-	-0.112	-0.180	-0.284	0.159	-0.569
	200	-	0.093	-0.196	-0.299	2.402	1.869
	500	-	-0.241	0.004	1.051	0.914	1.537
	1000	-	0.078	-0.161	-0.285	-0.035	-0.329
	2000	-	-0.257	-0.106	-0.106	0.341	0.254
	5000	-	0.035	0.132	-0.112	0.087	0.326
	10000	-	38.500	-56.500	71.000	-91.000	-124.500
	Mean	-	4.761	-7.108	7.912	-8.293	-12.056

Note. Mean relative error of *t* = -3.357; *RMSE* = 0.235 .

Table 31. Observed Probability of Type I Error under Independent Samples *t* Test with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.781	-	-	-	.401	.266
	20	.901	-	-	.858	.688	.860
	50	.371	.375	.309	.468	.439	.342
	100	.532	.579	.608	.654	.469	.788
	200	.915	.907	.932	.940	.717	.759
	500	.797	.845	.797	.598	.623	.515
	1000	.481	.447	.554	.614	.496	.635
	2000	.756	.817	.781	.781	.676	.696
	5000	.264	.248	.206	.321	.225	.139
	10000	.998	.937	.911	.885	.857	.805
	Mean	.680	.644	.637	.680	.559	.581

Table 32. Observed Test Statistic and its Relative Error under Independent Samples *t* Test with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
t	10	-0.288	-	-	-	-0.175	-0.395
	20	-0.127	-	-	-0.238	-0.384	-0.528
	50	0.902	0.868	0.858	0.790	0.771	0.459
	100	0.627	0.511	0.485	0.489	0.559	0.479
	200	-0.107	-0.131	-0.072	-0.161	-0.155	-0.353
	500	-0.257	-0.258	-0.207	-0.397	-0.369	-0.373
	1000	0.706	0.765	0.745	0.564	0.758	0.735
	2000	-0.311	-0.299	-0.386	-0.358	-0.338	-0.557
	5000	-1.117	-1.170	-1.140	-1.231	-1.249	-1.156
	10000	0.002	0.035	0.010	0.149	0.053	-0.007
	Mean	0.003	0.040	0.037	-0.044	-0.053	-0.170
Relative Error of *t*	10	-	-	-	-	-0.392	0.372
	20	-	-	-	0.874	2.024	3.157
	50	-	-0.038	-0.049	-0.124	-0.145	-0.491
	100	-	-0.185	-0.226	-0.220	-0.108	-0.236
	200	-	0.224	-0.327	0.505	0.449	2.299
	500	-	0.004	-0.195	0.545	0.436	0.451
	1000	-	0.084	0.055	-0.201	0.074	0.041
	2000	-	-0.039	0.241	0.151	0.087	0.791
	5000	-	0.047	0.021	0.102	0.118	0.035
	10000	-	16.500	4.000	73.500	25.500	-4.500
	Mean	-	2.075	0.440	8.348	2.804	0.192

Note. Mean relative error of $t = 2.782$; $RMSE = 0.137$.

Table 33. Observed Probability of Type I Error under Independent Samples *t* Test with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.781	-	-	-	.865	.703
	20	.900	-	-	.815	.705	.604
	50	.372	.390	.395	.433	.444	.648
	100	.532	.611	.629	.626	.577	.633
	200	.915	.896	.943	.872	.877	.724
	500	.797	.797	.836	.692	.712	.709
	1000	.480	.444	.456	.573	.449	.463
	2000	.756	.765	.700	.720	.735	.578
	5000	.264	.242	.254	.218	.212	.248
	10000	.998	.972	.992	.882	.958	.994
	Mean	.680	.640	.651	.648	.653	.630

Table 34. Power of the Test to Detect Medium Effect Size under Independent Samples *t* Test for Various Missing Data Handling Methods.

		Percentage of Missing Data					
Method	*n*	0%	1%	2%	5%	10%	20%
Listwise Deletion	10	.105	-	-	-	.094	.090
	20	.184	-	-	.174	.170	.152
	50	.410	.403	.394	.389	.374	.337
	100	.697	.692	.688	.673	.649	.598
	200	.940	.938	.936	.928	.916	.881
	500	1.000	1.000	1.000	1.000	1.000	.999
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000
Mean Imputation Regression Imputation EM Imputation Multiple imputation	10	.105	-	-	-	.105	.105
	20	.184	-	-	.184	.184	.184
	50	.410	.410	.410	.410	.410	.410
	100	.697	.697	.697	.697	.697	.697
	200	.940	.940	.940	.940	.940	.940
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000

Note. EM = Expectation Maximization. Medium effect size defined as Cohen's $d = .5$. Power calculations assume a two-tailed test with $\alpha = .05$

Table 35. Observed Test Statistic and its Relative Error under Two Factor ANOVA with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	3.589	-	-	-	3.971	3.214
	20	5.855	-	-	5.430	5.463	2.899
	50	29.000	26.600	27.995	25.416	24.971	29.092
	100	65.271	65.751	64.176	62.527	61.297	42.704
	200	143.940	142.570	140.614	133.282	130.125	108.360
	500	336.291	330.677	327.825	309.401	310.362	261.333
	1000	561.629	558.090	547.609	533.406	507.345	464.539
	2000	1283.276	1268.889	1257.975	1215.334	1163.336	1025.496
	5000	2768.598	2753.676	2714.183	2601.131	2461.299	2260.333
	10000	5624.048	5569.528	5519.549	5347.669	4992.234	4407.322
	Mean	1082.150	1339.473	1324.991	1137.066	966.040	860.529
Relative Error of *F*	10	-	-	-	-	0.106	-0.104
	20	-	-	-	-0.073	-0.067	-0.505
	50	-	-0.083	-0.035	-0.124	-0.139	0.003
	100	-	0.007	-0.017	-0.042	-0.061	-0.346
	200	-	-0.010	-0.023	-0.074	-0.096	-0.247
	500	-	-0.017	-0.025	-0.080	-0.077	-0.223
	1000	-	-0.006	-0.025	-0.050	-0.097	-0.173
	2000	-	-0.011	-0.020	-0.053	-0.093	-0.201
	5000	-	-0.005	-0.020	-0.060	-0.111	-0.184
	10000	-	-0.010	-0.019	-0.049	-0.112	-0.216
	Mean	-	-0.017	-0.023	-0.067	-0.075	-0.220

Note. Mean relative error of F = -0.086; *RMSE* = 233.944.

Table 36. Observed Probability of Type I Error under Two Factor ANOVA with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.086	-	-	-	.086	.144
	20	.005	-	-	.007	.008	**.079**
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.009	< .001	< .001	.001	.009	.022

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples. The *p* values for imputed data samples that generate contradictory test results at α =.05 relative to complete samples are in shown in boldface.

Table 37. Observed Test Statistic and its Relative Error under Two Factor ANOVA with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	3.589	-	-	-	4.618	4.638
	20	5.855	-	-	5.772	6.251	1.678
	50	29.000	22.965	29.176	19.771	15.336	28.167
	100	65.271	58.737	51.299	54.024	61.910	21.436
	200	143.940	135.837	135.778	116.979	95.072	56.598
	500	336.291	310.954	297.016	252.823	229.060	147.963
	1000	561.629	544.046	514.682	471.233	389.377	316.803
	2000	1283.276	1215.769	1175.372	1062.819	909.201	603.494
	5000	2768.598	2702.371	2597.715	2255.360	1896.769	1522.330
	10000	5624.048	5433.794	5240.918	4744.653	3860.974	2804.578
	Mean	1082.150	1303.059	1255.245	998.159	746.857	550.769
Relative Error of *F*	10	-	-	-	-	0.287	0.292
	20	-	-	-	-0.014	0.068	-0.713
	50	-	-0.208	0.006	-0.318	-0.471	-0.029
	100	-	-0.100	-0.214	-0.172	-0.051	-0.672
	200	-	-0.056	-0.057	-0.187	-0.340	-0.607
	500	-	-0.075	-0.117	-0.248	-0.319	-0.560
	1000	-	-0.031	-0.084	-0.161	-0.307	-0.436
	2000	-	-0.053	-0.084	-0.172	-0.292	-0.530
	5000	-	-0.024	-0.062	-0.185	-0.315	-0.450
	10000	-	-0.034	-0.068	-0.156	-0.313	-0.501
	Mean	-	-0.073	-0.085	-0.179	-0.205	-0.421

Note. Mean relative error of F =-0.203; *RMSE* = 586.159 .

Table 38. Observed Probability of Type I Error under Two Factor ANOVA with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.086	-	-	-	.053	.053
	20	.005	-	-	.005	.004	**.207**
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.009	< .001	< .001	.001	.006	.026

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples. The *p* values for imputed data samples that generate contradictory test results at α =.05 relative to complete samples are in shown in boldface.

Table 39. Observed Test Statistic and its Relative Error under Two Factor ANOVA with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	3.589	-	-	-	3.715	2.989
	20	5.855	-	-	5.574	4.883	1.746
	50	29.000	28.998	27.039	17.789	12.361	21.805
	100	65.271	61.191	64.235	59.505	42.002	47.635
	200	143.940	134.879	137.236	136.642	102.296	73.465
	500	336.291	319.964	318.870	283.246	297.603	238.662
	1000	561.629	553.229	532.466	487.230	401.849	349.031
	2000	1283.276	1248.641	1246.706	1114.854	1003.149	744.278
	5000	2768.598	2672.755	2655.559	2337.022	2156.150	1743.464
	10000	5624.048	5476.958	5348.032	4935.868	4384.446	3430.846
	Mean	1082.150	1312.077	1291.268	1041.970	840.845	665.392
Relative Error of *F*	10	-	-	-	-	0.035	-0.167
	20	-	-	-	-0.048	-0.166	-0.702
	50	-	0.000	-0.068	-0.387	-0.574	-0.248
	100	-	-0.063	-0.016	-0.088	-0.356	-0.270
	200	-	-0.063	-0.047	-0.051	-0.289	-0.490
	500	-	-0.049	-0.052	-0.158	-0.115	-0.290
	1000	-	-0.015	-0.052	-0.132	-0.284	-0.379
	2000	-	-0.027	-0.028	-0.131	-0.218	-0.420
	5000	-	-0.035	-0.041	-0.156	-0.221	-0.370
	10000	-	-0.026	-0.049	-0.122	-0.220	-0.390
	Mean	-	-0.035	-0.044	-0.141	-0.241	-0.373

Note. Mean relative error of F = -0.179; *RMSE* = 448.507 .

Table 40. Observed Probability of Type I Error under Two Factor ANOVA with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.086	-	-	-	.080	.118
	20	.005	-	-	.006	.010	**.192**
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.009	< .001	< .001	.001	.009	.031

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples. The *p* values for imputed data samples that generate contradictory test results at α =.05 relative to complete samples are in shown in boldface.

Table 41. Observed Test Statistic and its Relative Error under Two Factor ANOVA with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	3.589	-	-	-	4.796	4.160
	20	5.855	-	-	5.828	5.341	2.217
	50	29.000	28.978	27.905	25.051	23.151	28.336
	100	65.271	68.381	65.732	61.363	49.892	53.684
	200	143.940	136.195	140.236	139.678	116.339	92.207
	500	336.291	321.303	331.877	311.838	312.733	280.173
	1000	561.629	554.133	543.864	529.610	505.576	454.144
	2000	1283.276	1263.257	1261.092	1206.178	1148.653	929.951
	5000	2768.598	2735.257	2700.927	2529.402	2413.326	2244.215
	10000	5624.048	5551.572	5507.043	5289.489	4857.671	4357.438
	Mean	1082.150	1332.385	1322.335	1122.049	943.748	844.653
Relative Error of *F*	10	-	-	-	-	0.336	0.159
	20	-	-	-	-0.005	-0.088	-0.621
	50	-	-0.001	-0.038	-0.136	-0.202	-0.023
	100	-	0.048	0.007	-0.060	-0.236	-0.178
	200	-	-0.054	-0.026	-0.030	-0.192	-0.359
	500	-	-0.045	-0.013	-0.073	-0.070	-0.167
	1000	-	-0.013	-0.032	-0.057	-0.100	-0.191
	2000	-	-0.016	-0.017	-0.060	-0.105	-0.275
	5000	-	-0.012	-0.024	-0.086	-0.128	-0.189
	10000	-	-0.013	-0.021	-0.059	-0.136	-0.225
	Mean	-	-0.013	-0.020	-0.063	-0.092	-0.207

Note. Mean relative error of F = -0.085; *RMSE* = 256.517.

Table 42. Observed Probability of Type I Error under Two Factor ANOVA with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.086	-	-	-	**.049**	.065
	20	.005	-	-	.005	.007	.116
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.009	< .001	< .001	.001	.006	.018

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples. The *p* values for imputed data samples that generate contradictory test results at α =.05 relative to complete samples are in shown in boldface.

Table 43. Observed Test Statistic and its Relative Error under Two Factor ANOVA with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	3.589	-	-	-	3.160	3.099
	20	5.855	-	-	5.590	5.458	6.687
	50	29.000	28.346	28.578	26.207	29.574	32.549
	100	65.271	68.688	67.568	64.819	59.984	60.610
	200	143.940	144.256	143.687	140.869	132.973	136.572
	500	336.291	334.854	339.760	333.866	347.102	316.559
	1000	561.629	562.357	560.838	567.313	575.468	571.461
	2000	1283.276	1283.495	1288.920	1278.800	1282.219	1296.840
	5000	2768.598	2775.462	2766.390	2740.074	2771.833	2795.155
	10000	5624.048	5614.683	5640.507	5627.153	5576.242	5567.529
	Mean	1082.150	1351.518	1354.531	1198.299	1078.401	1078.706
Relative Error of *F*	10	-	-	-	-	-0.120	-0.137
	20	-	-	-	-0.045	-0.068	0.142
	50	-	-0.023	-0.015	-0.096	0.020	0.122
	100	-	0.052	0.035	-0.007	-0.081	-0.071
	200	-	0.002	-0.002	-0.021	-0.076	-0.051
	500	-	-0.004	0.010	-0.007	0.032	-0.059
	1000	-	0.001	-0.001	0.010	0.025	0.018
	2000	-	0.000	0.004	-0.003	-0.001	0.011
	5000	-	0.002	-0.001	-0.010	0.001	0.010
	10000	-	-0.002	0.003	0.001	-0.009	-0.010
	Mean	-	0.004	0.004	-0.020	-0.028	-0.003

Note. Mean relative error of F = -0.009; $RMSE$ = 13.975 .

Table 44. Observed Probability of Type I Error under Two Factor ANOVA with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.086	-	-	-	.107	.111
	20	.079	-	-	**.006**	**.006**	**.003**
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.017	< .001	< .001	.001	.011	.011

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples. The *p* values for imputed data samples that generate contradictory test results at α =.05 relative to complete samples are in shown in boldface.

Table 45. Power of the Test to Detect Medium Effect Size under Two Factor ANOVA for Various Missing Data Handling Methods.

		Percentage of Missing Data					
Method	*n*	0%	1%	2%	5%	10%	20%
Listwise Deletion	10	.061	-	-	-	.058	.055
	20	.092	-	-	.089	.086	.079
	50	.207	.203	.198	.194	.186	.166
	100	.426	.421	.417	.404	.382	.338
	200	.775	.770	.765	.749	.721	.659
	500	.996	.996	.996	.995	.992	.984
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000
Mean Imputation Regression Imputation EM Imputation Multiple imputation	10	.061	-	-	-	.061	.061
	20	.092	-	-	.092	.092	.092
	50	.207	.207	.207	.207	.207	.207
	100	.426	.426	.426	.426	.426	.426
	200	.775	.775	.775	.775	.775	.775
	500	.996	.996	.996	.996	.996	.996
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000

Note. Medium effect size defined as $f = .25$. Power calculations are based on $\alpha = .05$, numerator $df = 5$, and six groups.

Table 46. Observed Test Statistic and its Relative Error under Multiple Regression with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	1.941	-	-	-	1.887	2.043
	20	2.162	-	-	1.992	1.979	.466
	50	24.052	21.472	23.406	22.240	20.685	25.945
	100	72.197	72.345	70.934	67.691	75.231	44.431
	200	100.474	100.330	98.808	93.497	95.801	85.395
	500	237.166	234.137	224.716	215.746	202.375	184.442
	1000	413.439	414.606	403.868	386.969	359.771	313.525
	2000	940.298	929.037	912.588	892.898	831.552	793.092
	5000	2234.596	2211.882	2198.128	2133.023	1983.055	1786.753
	10000	4409.374	4357.618	4327.396	4180.389	3994.858	3463.748
	Mean	843.570	1042.678	1032.481	888.272	756.719	669.984
Relative Error of *F*	10	-	-	-	-	-0.028	0.053
	20	-	-	-	-0.079	-0.085	-0.784
	50	-	-0.107	-0.027	-0.075	-0.140	0.079
	100	-	0.002	-0.017	-0.062	0.042	-0.385
	200	-	-0.001	-0.017	-0.069	-0.047	-0.150
	500	-	-0.013	-0.052	-0.090	-0.147	-0.222
	1000	-	0.003	-0.023	-0.064	-0.130	-0.242
	2000	-	-0.012	-0.029	-0.050	-0.116	-0.157
	5000	-	-0.010	-0.016	-0.045	-0.113	-0.200
	10000	-	-0.012	-0.019	-0.052	-0.094	-0.214
	Mean	-	-0.019	-0.025	-0.065	-0.086	-0.222

Note. Mean relative error of F = -0.089; *RMSE* =180.117 .

Table 47. Observed Probability of Type I Error under Multiple Regression with Listwise Deletion.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.242	-	-	-	.277	.292
	20	.123	-	-	.151	.157	.760
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.037	< .001	< .001	.017	.043	.105

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples.

Table 48. Observed Test Statistic and its Relative Error under Multiple Regression with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	1.941	-	-	-	2.293	2.635
	20	2.162	-	-	2.125	1.861	0.504
	50	24.052	17.619	22.341	21.471	20.545	23.320
	100	72.197	65.282	62.756	49.610	63.206	24.327
	200	100.474	97.655	97.891	85.076	91.005	60.597
	500	237.166	228.253	202.951	190.395	160.836	130.566
	1000	413.439	406.685	394.962	367.084	319.130	228.919
	2000	940.298	914.827	881.712	809.967	685.104	586.843
	5000	2234.596	2171.348	2123.251	1944.931	1650.359	1341.088
	10000	4409.374	4260.272	4174.108	3842.797	3386.457	2531.033
	Mean	843.570	1020.243	994.997	812.606	638.080	492.983
Relative Error of *F*	10	-	-	-	-	0.181	0.358
	20	-	-	-	-0.017	-0.139	-0.767
	50	-	-0.267	-0.071	-0.107	-0.146	-0.030
	100	-	-0.096	-0.131	-0.313	-0.125	-0.663
	200	-	-0.028	-0.026	-0.153	-0.094	-0.397
	500	-	-0.038	-0.144	-0.197	-0.322	-0.449
	1000	-	-0.016	-0.045	-0.112	-0.228	-0.446
	2000	-	-0.027	-0.062	-0.139	-0.271	-0.376
	5000	-	-0.028	-0.050	-0.130	-0.261	-0.400
	10000	-	-0.034	-0.053	-0.128	-0.232	-0.426
	Mean	-	-0.067	-0.073	-0.144	-0.164	-0.360

Note. Mean relative error of F = -0.170; *RMSE* = 379.923 .

Table 49. Observed Probability of Type I Error under Multiple Regression with Mean Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.242	-	-	-	.194	.158
	20	.123	-	-	.128	.170	.733
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.037	< .001	< .001	.014	.036	.089

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples.

Table 50. Observed Test Statistic and its Relative Error under Multiple Regression with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	1.941	-	-	-	2.258	3.302
	20	2.162	-	-	2.193	2.868	0.891
	50	24.052	24.241	24.000	22.184	17.972	30.068
	100	72.197	75.875	75.118	80.056	80.804	63.820
	200	100.474	103.102	99.089	99.500	103.128	110.289
	500	237.166	237.605	228.305	234.901	222.572	242.004
	1000	413.439	422.090	412.351	403.930	385.499	408.363
	2000	940.298	941.440	936.435	935.540	916.736	967.085
	5000	2234.596	2241.756	2239.331	2224.171	2200.772	2238.061
	10000	4409.374	4399.125	4422.547	4394.603	4481.907	4363.537
	Mean	843.570	1055.654	1054.647	933.009	841.452	842.742
Relative Error of *F*	10	-	-	-	-	0.163	0.701
	20	-	-	-	0.014	0.327	-0.588
	50	-	0.008	-0.002	-0.078	-0.253	0.250
	100	-	0.051	0.040	0.109	0.119	-0.116
	200	-	0.026	-0.014	-0.010	0.026	0.098
	500	-	0.002	-0.037	-0.010	-0.062	0.020
	1000	-	0.021	-0.003	-0.023	-0.068	-0.012
	2000	-	0.001	-0.004	-0.005	-0.025	0.028
	5000	-	0.003	0.002	-0.005	-0.015	0.002
	10000	-	-0.002	0.003	-0.003	0.016	-0.010
	Mean	-	0.014	-0.002	-0.001	0.023	0.037

Note. Mean relative error of $F = 0.015$; $RMSE = 16.482$.

Table 51. Observed Probability of Type I Error under Multiple Regression with Regression Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.242	-	-	-	.198	.111
	20	.123	-	-	.119	.060	.493
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.037	< .001	< .001	.013	.026	.060

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples.

Table 52. Observed Test Statistic and its Relative Error under Multiple Regression with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	1.941	-	-	-	2.439	3.948
	20	2.162	-	-	2.142	2.654	0.837
	50	24.052	23.916	25.367	24.738	24.254	37.065
	100	72.197	75.317	75.242	79.377	90.676	78.413
	200	100.474	102.626	101.931	103.79	112.321	133.358
	500	237.166	239.570	236.763	242.813	257.191	293.496
	1000	413.439	423.386	418.901	424.369	438.811	502.068
	2000	940.298	946.860	949.538	992.616	1037.336	1220.967
	5000	2234.596	2257.499	2288.334	2369.415	2470.648	2767.297
	10000	4409.374	4454.529	4508.127	4631.143	4935.446	5458.975
	Mean	843.570	1065.463	1075.525	985.600	937.178	1049.642
Relative Error of *F*	10	-	-	-	-	0.257	1.034
	20	-	-	-	-0.009	0.228	-0.613
	50	-	-0.006	0.055	0.029	0.008	0.541
	100	-	0.043	0.042	0.099	0.256	0.086
	200	-	0.021	0.015	0.033	0.118	0.327
	500	-	0.010	-0.002	0.024	0.084	0.238
	1000	-	0.024	0.013	0.026	0.061	0.214
	2000	-	0.007	0.010	0.056	0.103	0.298
	5000	-	0.010	0.024	0.060	0.106	0.238
	10000	-	0.010	0.022	0.050	0.119	0.238
	Mean	-	0.015	0.022	0.041	0.134	0.260

Note. Mean relative error of $F = 0.102$; $RMSE = 205.791$.

Table 53. Observed Probability of Type I Error under Multiple Regression with Expectation Maximization Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.242	-	-	-	.177	.082
	20	.123	-	-	.126	.074	.523
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.037	< .001	< .001	.014	.025	.061

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples.

Table 54. Observed Test Statistic and its Relative Error under Multiple Regression with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
F	10	1.941	-	-	-	1.791	2.201
	20	2.162	-	-	2.105	2.288	1.717
	50	24.052	23.146	22.947	24.197	22.671	27.543
	100	72.197	75.460	75.732	72.241	71.994	62.219
	200	100.474	99.859	100.955	101.314	100.372	100.076
	500	237.166	236.042	232.483	229.083	230.211	233.884
	1000	413.439	417.186	411.013	406.268	404.400	397.310
	2000	940.298	938.271	938.118	947.741	939.846	980.626
	5000	2234.596	2227.696	2240.287	2239.891	2212.101	2233.758
	10000	4409.374	4402.731	4418.000	4406.738	4455.345	4400.927
	Mean	843.570	1052.549	1054.942	936.620	844.102	844.026
Relative Error of *F*	10	-	-	-	-	-0.077	0.134
	20	-	-	-	-0.026	0.058	-0.206
	50	-	-0.038	-0.046	0.006	-0.057	0.145
	100	-	0.045	0.049	0.001	-0.003	-0.138
	200	-	-0.006	0.005	0.008	-0.001	-0.004
	500	-	-0.005	-0.020	-0.034	-0.029	-0.014
	1000	-	0.009	-0.006	-0.017	-0.022	-0.039
	2000	-	-0.002	-0.002	0.008	0.000	0.043
	5000	-	-0.003	0.003	0.002	-0.010	0.000
	10000	-	-0.002	0.002	-0.001	0.010	-0.002
	Mean	-	0.000	-0.002	-0.006	-0.013	-0.008

Note. Mean relative error of F = -0.006; $RMSE$ = 10.870 .

Table 55. Observed Probability of Type I Error under Multiple Regression with Multiple Imputation.

		Percentage of Missing Data					
	n	0%	1%	2%	5%	10%	20%
p	10	.242	-	-		0.268	0.205
	20	.123	-	-	0.131	0.108	0.199
	50	< .001	< .001	< .001	< .001	< .001	< .001
	100	< .001	< .001	< .001	< .001	< .001	< .001
	200	< .001	< .001	< .001	< .001	< .001	< .001
	500	< .001	< .001	< .001	< .001	< .001	< .001
	1000	< .001	< .001	< .001	< .001	< .001	< .001
	2000	< .001	< .001	< .001	< .001	< .001	< .001
	5000	< .001	< .001	< .001	< .001	< .001	< .001
	10000	< .001	< .001	< .001	< .001	< .001	< .001
	Mean	.037	< .001	< .001	.015	.038	.040

Note. Shaded cells highlight the tendency of *p* values to behave differently in small samples.

Table 56. Power of the Test to Detect Medium Effect Size under Multiple Regression for Various Missing Data Handling Methods.

		Percentage of Missing Data					
Method	*n*	0%	1%	2%	5%	10%	20%
Listwise Deletion	10	.088	-	-	-	.079	.071
	20	.188	-	-	.177	.167	.146
	50	.524	.513	.503	.492	.471	.415
	100	.874	.870	.866	.853	.829	.773
	200	.997	.996	.996	.995	.992	.984
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000
Mean Imputation Regression Imputation EM Imputation Multiple imputation	10	.088	.088	.088	.088	.088	.088
	20	.188	.188	.188	.188	.188	.188
	50	.524	.524	.524	.524	.524	.524
	100	.874	.874	.874	.874	.874	.874
	200	.997	.997	.997	.997	.997	.997
	500	1.000	1.000	1.000	1.000	1.000	1.000
	1000	1.000	1.000	1.000	1.000	1.000	1.000
	2000	1.000	1.000	1.000	1.000	1.000	1.000
	5000	1.000	1.000	1.000	1.000	1.000	1.000
	10000	1.000	1.000	1.000	1.000	1.000	1.000

Note. EM = Expectation Maximization. Medium effect size defined as $f^2 = .15$. Power calculations assume $\alpha = .05$ and one set of four predictors in the linear regression equation.

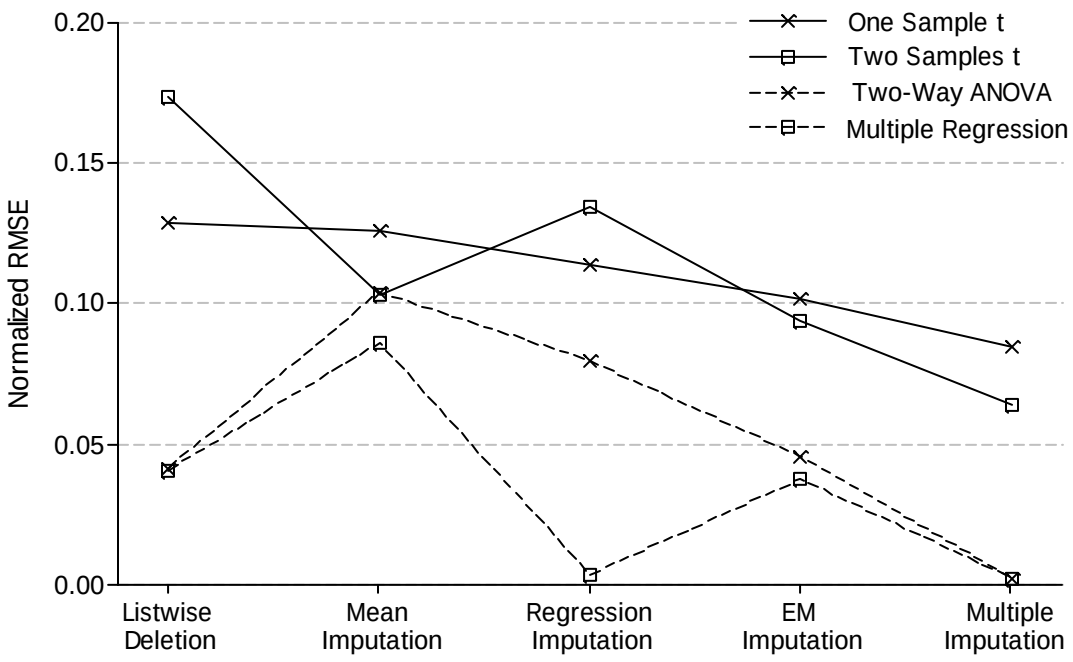
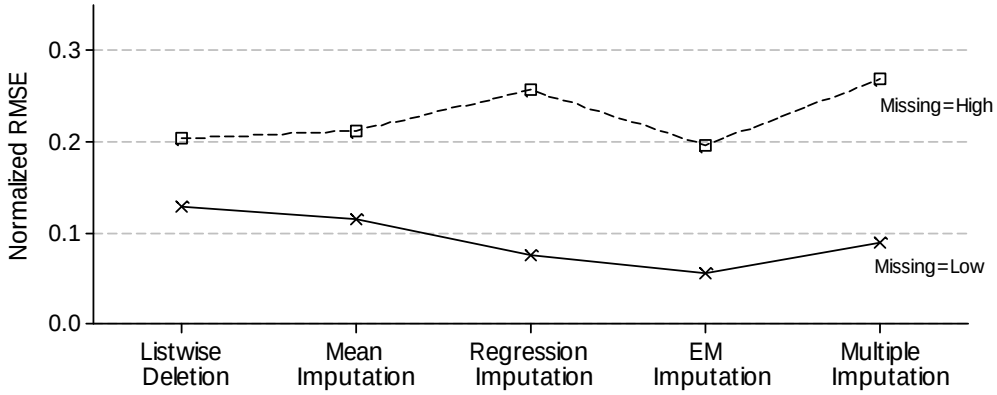
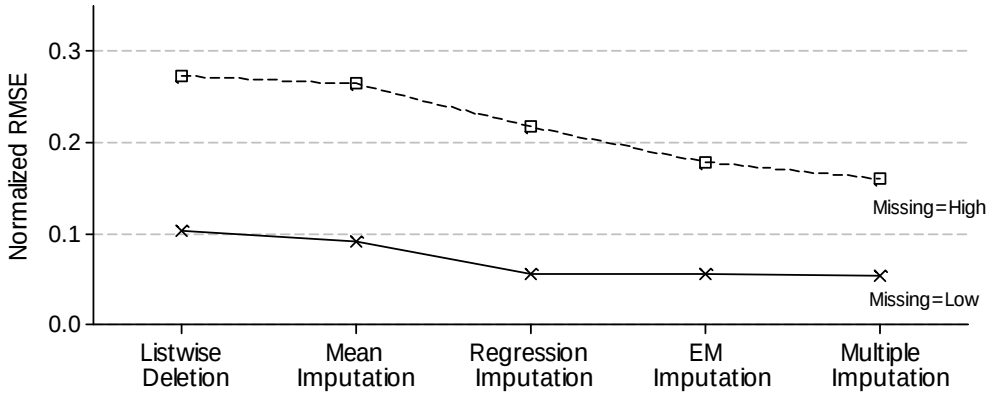


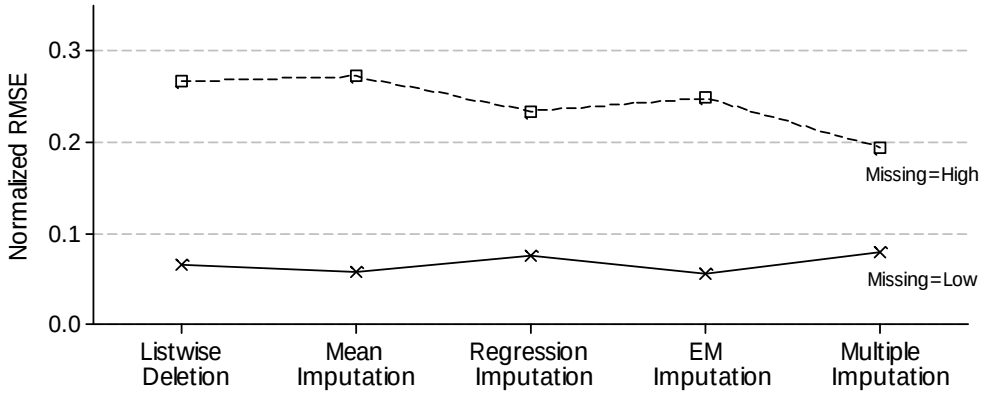
Figure 24. The average effect of missing data handling method on accuracy of estimation for various methods of analysis.



a. Sample size = Small

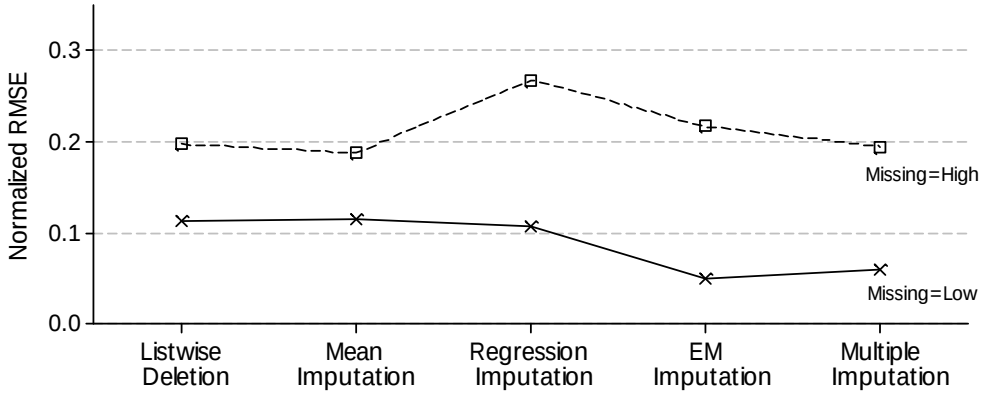


b. Sample size = Medium

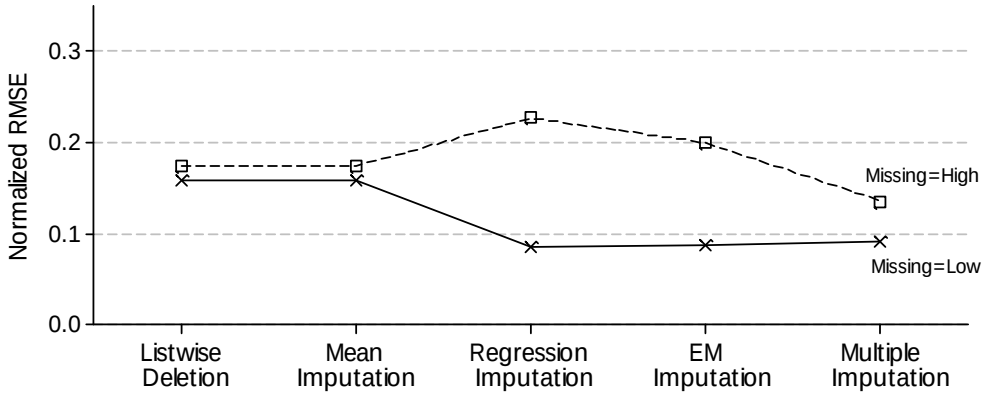


c. Sample size = Large

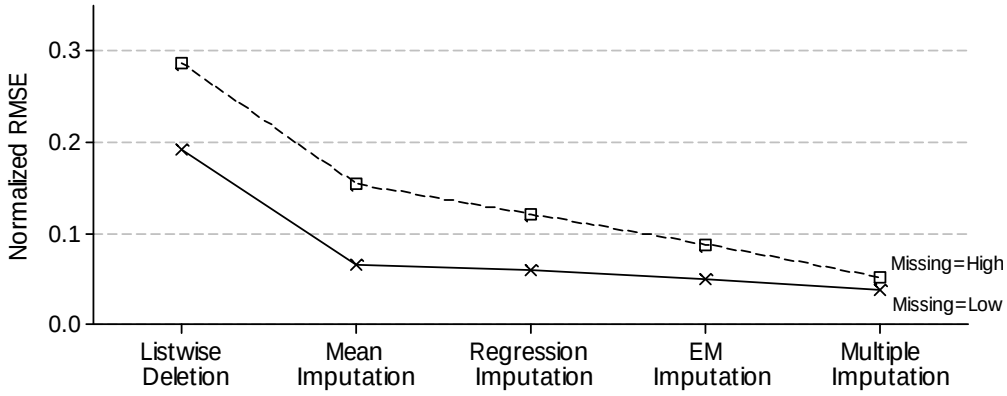
Figure 25. Normalized root mean squared error plotted for the five missing data handling methods by incidence of missing data (Low, High) and sample size (Small, Medium, Large) when the method of analysis is one sample *t* test.



a. Sample size = Small

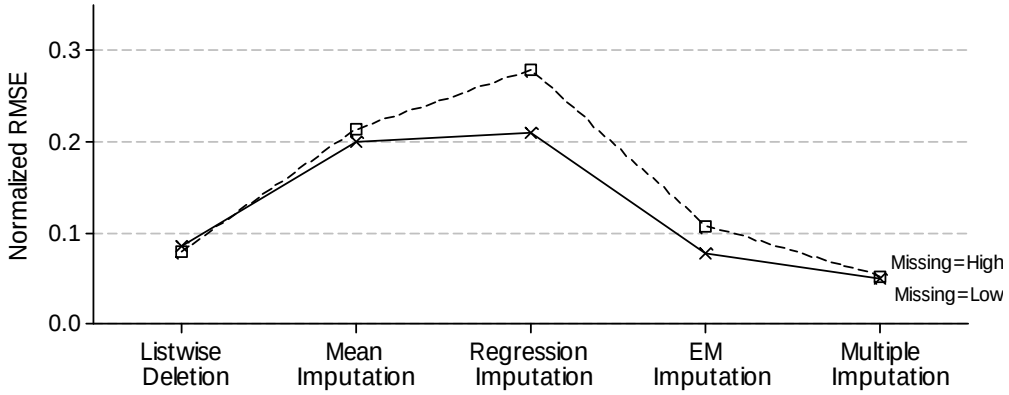


b. Sample size = Medium

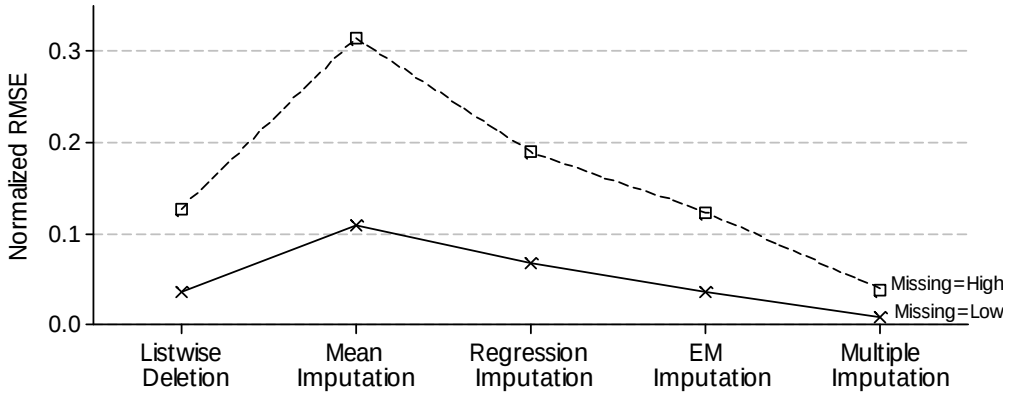


c. Sample size = Large

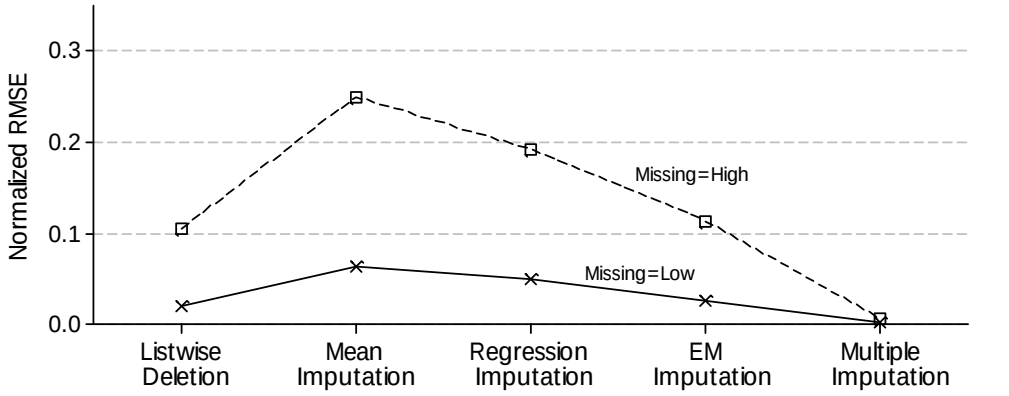
Figure 26. Normalized root mean squared error plotted for the five missing data handling methods by incidence of missing data (Low, High) and sample size (Small, Medium, Large) when the method of analysis is independent samples *t* test.



a. Sample size = Small

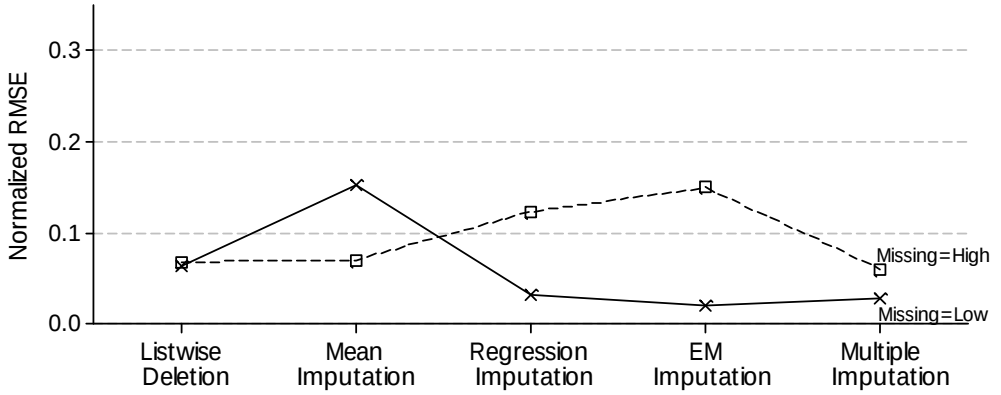


b. Sample size = Medium

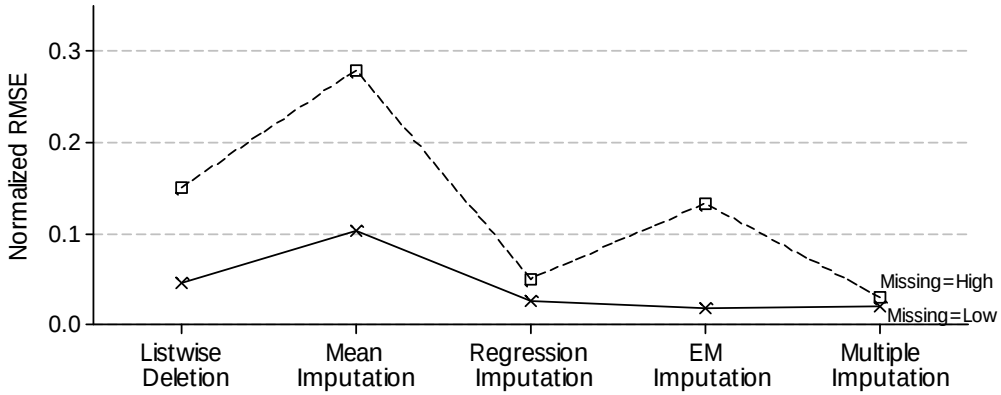


c. Sample size = Large

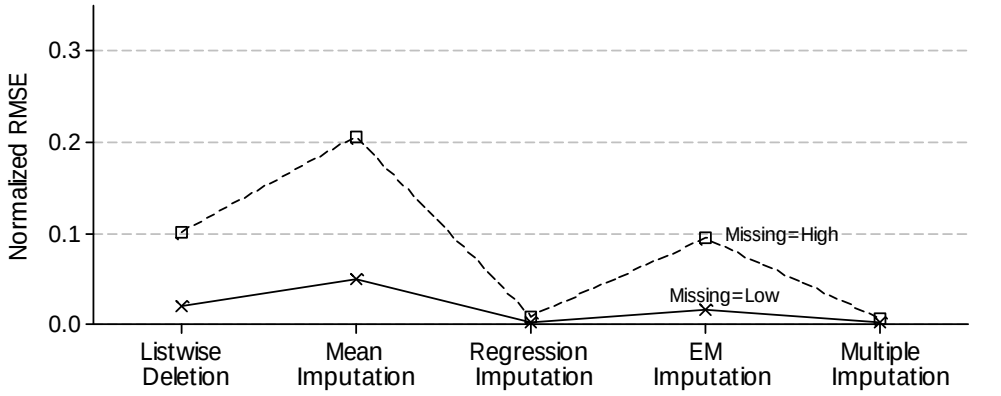
Figure 27. Normalized root mean squared error plotted for the five missing data handling methods by incidence of missing data (Low, High) and sample size (Small, Medium, Large) when the method of analysis is two-way ANOVA.



a. Sample size = Small



b. Sample size = Medium



c. Sample size = Large

Figure 28. Normalized root mean squared error plotted for the five missing data handling methods by incidence of missing data (Low, High) and sample size (Small, Medium, Large) when the method of analysis is multiple regression.

Figure 24 plots normalized RMSE as a function of missing data handling method with a separate line for each method of analysis. The normalized RMSE presented in this graph is averaged over both sample size and proportion of missing data. It can be observed from this plot that multiple imputation is the best missing data handling method because it produces 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 perform better than listwise deletion although the difference between listwise deletion and mean imputation is small. For independent samples *t* test, listwise deletion does not perform very well but mean imputation does. Furthermore, for independent samples *t* test, the performance of mean imputation and EM imputation is almost the same. For two-way ANOVA, listwise deletion is as good as EM imputation and better than regression imputation and mean imputation, the latter being the most error-prone method. For multiple regression, regression imputation works almost as well as multiple imputation which produces the smallest normalized RMSE, listwise deletion and EM imputation behave similarly, and mean imputation is clearly inferior to all other missing data handling methods. The reason why regression imputation performs so well when the analytical method is multiple regression is 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 24 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 24 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 are presented in Figures 25 through 28. Figure 25 shows that for one sample *t* test: for small samples ($n \leq 50$), EM imputation works best whether the proportion of missing data is low ($m \leq 5\%$) or high ($m > 5\%$.); for medium samples ($50 < n < 1{,}000$), multiple imputation works best regardless of proportion of missing data; and for large samples ($n \geq 1{,}000$), EM imputation works best when proportion of missing data is low and multiple imputation works best when proportion of missing data is high. It should be noted here that even though we have identified the best missing data 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. For example, in panel *b* of Figure 25, although multiple imputation performs best when proportion of missing data is small, the gain in performance compared to regression imputation or EM imputation is trivial. Findings for independent samples *t* test, two-way ANOVA, and multiple regression are similarly summarized in Figures 26, 27, and 28 respectively. They can be interpreted in the same way as Figure 25.

Power comparisons for the four methods of analysis are provided in Tables 23, 34, 45, and 56. The information provided in these tables suggests that with listwise

deletion and medium effect sizes as defined by Cohen (1992): one sample *t* test achieves 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 achieves power of .8 at sample sizes between 100 and 200 for any proportion of missing data ranging between 1% and 20%; 2×3 ANOVA achieves 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 achieves 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 is 20%. It should be noted that for the four imputation methods, power values at all sample sizes are exactly identical to those of the complete data because after imputation sample sizes are at their maximum.

Information presented in Tables 24 through 33, 35 through 44, and 46 through 55 is synthesized as a decision tree in Table 57. Out of the 24 possible situations listed in Table 57 based on various combinations of method of analysis (one sample t test, independent samples t test, two-way ANOVA, multiple regression), sample size (low, 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 EM, in one case (4.2%) the best method was regression imputation, and in 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

Table 57. 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]	EM	EM	MI	MI	EM	MI	EM	M	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	EM	MI	EM	MI	EM	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.

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.

Empirical Data Example 1 - Large Sample

In order to allow comparison with simulated data results, a multiple regression equation with three continuous predictors, reading achievement, math anxiety, and home educational resources, and one categorical predictor with two levels, gender based on PISA 2003 data, was estimated in order to explain variation in math achievement. The maximum sample size was 5,455 as one case had missing data on gender. Results for the full dataset and the datasets based on various missing data handling methods are presented in Tables 58 and 59 under low, m = 5% (n = 5,182) and high, m = 10% (n = 4,910) missing data conditions. 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 58 and 59 are identical. The only exception is where 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 58 and 59 suggest that regression imputation multiple imputation methods provide effect size estimates that closely match

Table 58. 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.

* $p < .05$; ** $p < .01$; *** $p < .001$

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.

Table 59. Predicting Math Achievement: Multiple Regression Results with 10% Missing Data under Various Missing Data Handling Methods using PISA 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.

* $p < .05$; ** $p < .01$; *** $p < .001$

Empirical Data Example 2 - Small Sample

For the small sample illustration, U.S. Census Bureau (2000) data was 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 maximum 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 60 and 61 under low, $m = 5\%$ ($n = 197$) and high, $m = 10\%$ ($n = 186$) missing data conditions. These results show that, in terms of effect size, best results are obtained with listwise deletion ($d = .26$) and EM 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 ($t = 2.19$, $p = .030$), EM 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 ($t = 1.82$, $p = .071$), EM imputation ($t = 1.79$, $p = .075$), and

Table 60. 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	.045	2.47	1.23	.25	.926
Listwise deletion	2.08*	194.88	.039	2.62	1.26	.26	.915
Mean imputation	2.19*	204.66	.030	2.62	1.20	.27	.926
Regression imputation	1.84	204.08	.067	2.25	1.22	.23	.926
EM imputation	2.09*	204.64	.038	2.50	1.20	.26	.926
Multiple imputation	1.87*	204.08	.038	2.59	1.24	.21	.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 61. Independent Samples *t* Test Results for Educational 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	123	19.66	11.61	63	17.28	6.23
Mean imputation	135	19.66	11.07	72	17.28	5.82
Regression imputation	135	19.46	11.32	72	16.83	6.10
EM imputation	135	19.59	11.08	72	17.48	5.84
Multiple imputation	135	19.74	11.55	72	17.37	6.64

	t test statistics						
	t	*df*	*p*	Δ*M*	*SE*(Δ*M*)	*d*	*Power*
Full data	2.02*	204.98	.045	2.47	1.23	.25	.926
Listwise deletion	1.82	183.57	.071	2.37	1.31	.23	.894
Mean imputation	2.02*	204.98	.044	2.37	1.17	.25	.926
Regression imputation	2.18*	204.96	.031	2.63	1.21	.27	.926
EM imputation	1.79	204.99	.075	2.11	1.18	.22	.926
Multiple imputation	1.87	202.46	.071	2.37	1.27	.21	.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*.
* $p < .05$; ** $p < .01$; *** $p < .001$

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.

5. 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, by using uniform empirical and simulated samples, and uniform rates of change in sample size and proportion of missing data. By keeping all of these factors constant, any observed differences in performance of missing data handling methods can more or less be attributed directly 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. For example, Roth (1994) 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. Although past 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 that can be used to form general rules

of thumbs which can inform about which missing data handling method is best under which circumstances. The deficiencies in past research range from using 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, to ignoring the power-related analyses and a comprehensive measurement of biases associated with missing data techniques. The present study is an early 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 many respects the present study has been able to confirm and support the findings of earlier research. 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 thus 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. EM 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 57 which show that EM 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 data points depicted in Figures 25, 26, 27, and 28. The sum of scores across missing data handling methods revealed the following ranking and total scores: multiple imputation, 80; expectation maximization, 59; listwise deletion, 41; regression imputation, 39; and mean imputation, 21. 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 those methods in the order of their mathematical sophistication, ranging from the most desirable, multiple imputation which offers most realistic modeling of random variation, to the least desirable, 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, consider panel *b* of Figure 27 which depicts estimation accuracy results for a medium sample under two-way ANOVA and where under the high missing data condition listwise deletion performs better than mean imputation but worse than multiple imputation. For mean imputation in this scenario any positive effect of missing data imputation is dominated by the negative effect of higher measurement error due to that imputation. On the other hand, the reverse is true for multiple imputation where the positive effect is missing data imputation dominates the negative effect of higher 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 such 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 for Large Samples

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 is still representative of the target population. In situations where non-missing data represents the population well, listwise deletion provides an attractive choice. Although, technically, sophisticated procedures like multiple imputation and EM imputation provide superior accuracy gains in parameter estimation, the magnitude of such gains may be too small to justify the additional statistical complexity involved. In cases where non-missing data does not adequately represent the target population and missing data imputation is unavoidable, we recommend advanced methods like multiple imputation and EM imputation under both low and high missing data conditions as these method offer the largest gains in accuracy of parameter estimation for most methods of analysis.

Recommendations for Small Samples

Our general recommendation for researchers using small samples, $n \leq 50$, is that they give a serious consideration to listwise deletion over missing data imputation when both of the following conditions hold, (1) non-missing data still represents the target population well, and (2) statistical power is not an issue. The reason for this recommendation is that the measurement error introduced with missing data imputation tends to be large in small samples due to the smaller number of observations available that form the basis of any imputation method, and the corresponding gain in efficiency tends to be small, usually less than 10% (see Table 57). If either of these two conditions

is not satisfied, then the researcher must resort to missing data imputation. For small samples, in terms of gain in accuracy of estimation, the best available methods of missing data imputation are EM imputation and multiple imputation. Although strictly speaking multiple imputation on average performs better than EM imputation in small samples, in the light of simulation results and empirical data examples that were presented earlier 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 an 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.

Recommendations for Medium Samples

For medium sample sizes that lie between 50 and 1,000, depending on the choice of analytical method, the researcher may chose to treat her sample as either small or large. For example, if the target statistical power, say for a medium effect size, is .8, then a sample size of 100 may be more than adequate for one method of analysis (such as multiple regression with four predictors, actual power = .874), but may be quite inadequate for a different method (such as two-way ANOVA with six groups, actual power = .426). Thus, it is up to the researcher to decide whether she wants to treat her sample as small or large. If the issue is still confusing then one of the safest strategies is to go with either multiple imputation or EM imputation, ideally employing both methods in order to minimize the negative effects of measurement error introduced with imputation of missing data.

Our recommendations for choice of missing data handling method are summarized in Figure 29. If the missing data are MCAR and the resulting sample after listwise deletion provides adequate power for tests of hypotheses, then listwise deletion

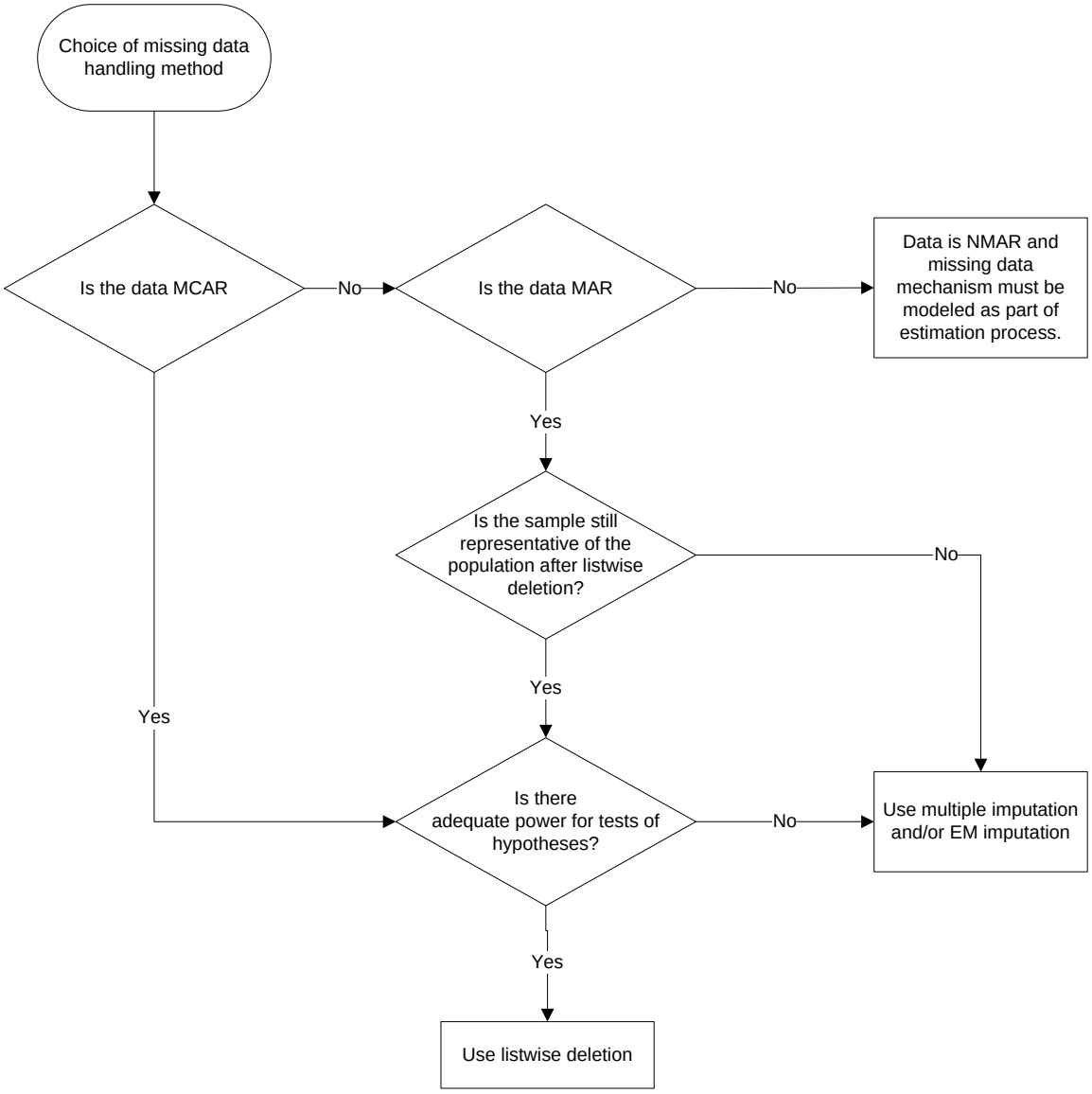


Figure 29. The decision process governing choice of missing data handling method.

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. Since 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). In Figure 29, we recommend multiple imputation and EM imputation as the methods of choice. Although in most instances the multiple imputation method is more efficient than EM imputation, we recommend using both of these methods in order to guard against the possibility of measurement error. introduced by such imputation. polluting the results of tests of hypotheses. This is especially important in scenarios involving small samples with a large proportion of missing data.

Scope for Future Research

There are several directions in which the future research can go in order to extend the work presented in this study. 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 can focus on simulation and investigate a larger subset of sample sizes, proportions of missing data, and missing data handling methods. Given the objectives of this study and its intended audience, only those missing data handling methods were investigated here that can be easily implemented with a specific software package, SPSS viz. listwise deletion, mean imputation, regression imputation, EM imputation, and multiple imputation. Future studies can expand this list of missing data handling methods and evaluate them under other statistical software programs common in the social sciences, such as SAS and Stata, in order to target a more mathematically sophisticated audience.

Finally, future studies can take an applied approach and use real-life datasets from various subfields of education 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.

Appendix

Illustration of Missing Data Imputation with SPSS

In order to demonstrate the imputation of missing data in SPSS/PASW 18 we employ state level data from the 2010 American Community Survey published by U.S. Census Bureau (2010). The data is provided in Table A. The steps described in the following paragraphs can be used to impute missing data in SPSS 18.0. This tutorial assumes a basic level of familiarity with SPSS, such as the ability to open and save a data file, enter data, use the point-and-click interface to browse through menu options etc. For those readers who are not familiar with these basic tasks we recommend various SPSS tutorials that are freely available on the world wide web (e.g. UCLA, 2010).

After opening the data file or entering the data manually from Table 62 into SPSS data editor, the program screen should look somewhat similar to the screenshot shown in Figure 30. This figure shows that some of the values for variables Income and PovertyRate are missing. Our task is to impute this missing data. With the exception of mean imputation that requires minor coding with SPSS syntax, the remaining imputation methods, regression imputation, EM imputation, and multiple imputation do not require any coding.

Table 62. Selected State-Level Data from the 2010 American Community Survey

ID	State	Abbr.	Population	MaleRatio	Edu.	Age	Fam. Size	Income	NoIns	PovRt	Grp.
1	Alabama	AL	4,785,298	0.94	21.9	37.8	3.16	21,993	14.6	**14.7**	1
2	Alaska	AK	713,985	1.09	27.9	33.8	3.24	30,598	19.9	7.2	3
3	Arizona	AZ	6,413,737	0.99	25.9	35.9	3.28	23,618	16.9	12.5	1
4	Arkansas	AR	2,921,606	0.96	19.5	37.3	3.09	20,725	17.5	14.1	1
5	California	CA	37,349,363	0.99	30.1	35.2	3.53	27,353	18.5	11.8	1
6	Colorado	CO	5,049,071	1	36.4	36	3.12	28,723	15.9	9.4	2
7	Connecticut	CT	3,577,073	0.95	35.5	40	3.15	35,078	9.1	**7.2**	1
8	Delaware	DE	899,769	0.94	27.8	38.8	3.24	27,729	9.7	8.1	1
9	District of Columbia	DC	604,453	0.9	50.1	33.9	3.37	41,240	7.6	14.1	1
10	Florida	FL	18,843,326	0.96	25.8	40.7	3.24	24,272	21.3	12	1
11	Georgia	GA	9,712,587	0.95	27.3	35.4	3.3	23,383	19.7	13.7	1
12	Hawaii	HI	1,363,621	1.01	29.5	38.7	3.59	27,537	7.9	7.4	3
13	Idaho	ID	1,571,450	1	24.4	34.7	3.2	20,991	17.7	11.6	2
14	Illinois	IL	12,843,166	0.96	30.7	36.6	3.28	27,325	13.8	10.1	1
15	Indiana	IN	6,490,621	0.97	22.7	36.9	3.11	22,806	14.8	11	1
16	Iowa	IA	3,049,883	0.98	25	38.2	2.97	**24,883**	9.3	8.2	1
17	Kansas	KS	2,859,169	0.98	29.8	36.2	3.12	24,911	13.9	9.5	1
18	Kentucky	KY	4,346,266	0.97	20.5	38	3.06	21,706	15.3	14.5	1
19	Louisiana	LA	4,544,228	0.96	21.4	35.9	3.21	22,862	17.8	14.5	1
20	Maine	ME	1,327,567	0.96	26.8	42.7	2.9	24,950	10.1	8.8	1
21	Maryland	MD	5,785,982	0.94	36.1	38	3.23	33,772	11.3	6.6	1
22	Massachusetts	MA	6,557,254	0.93	39	39.1	3.13	33,203	4.4	8.2	1
23	Michigan	MI	9,877,574	0.96	25.2	39	3.13	23,622	12.4	12.1	1
24	Minnesota	MN	5,310,584	0.98	31.8	37.5	3.06	28,563	9.1	7.5	1
25	Mississippi	MS	2,970,036	0.94	19.5	36.1	3.25	19,096	18.2	17.8	1
26	Missouri	MO	5,996,231	0.96	25.5	38	3.07	**23,920**	13.2	10.6	1
27	Montana	MT	990,898	1.02	28.8	40.1	2.96	23,552	17.3	10	3
28	Nebraska	NE	1,830,429	0.98	28.6	36.3	3.06	24,744	11.5	8.8	1
29	Nevada	NV	2,704,642	1.02	21.7	36.3	3.31	25,284	22.6	11.1	3
30	New Hampshire	NH	1,316,759	0.97	32.8	41.3	2.98	30,949	11.1	5.3	1
31	New Jersey	NJ	8,801,624	0.95	35.4	39	3.3	33,555	13.2	7.8	1
32	New Mexico	NM	2,065,932	0.98	25	36.7	3.28	22,150	19.6	**15.7**	1
33	New York	NY	19,392,283	0.94	32.6	38	3.26	30,011	11.9	11.5	1
34	North Carolina	NC	9,561,558	0.95	26.5	37.3	3.1	23,432	16.8	13.3	1
35	North Dakota	ND	674,499	1.04	27.6	37.4	2.92	26,021	9.8	7.8	3

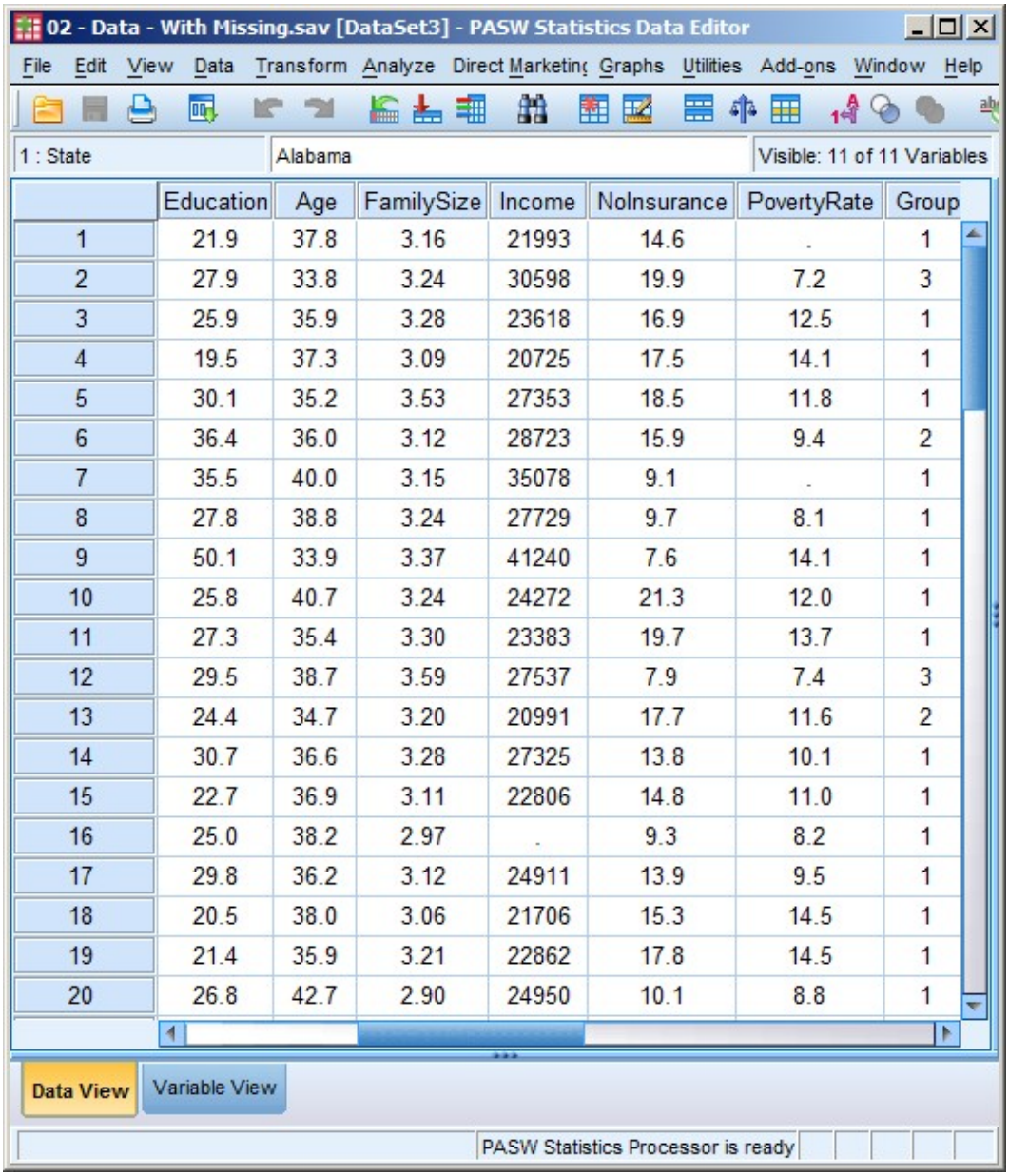
36 Ohio	OH	11,536,182	0.95	24.6	38.9	3.06	23,975	12.3	**11.8**	1
37 Oklahoma	OK	3,761,702	0.97	22.9	36.3	3.13	**22,254**	18.9	12.7	1
38 Oregon	OR	3,838,957	0.98	28.8	38.5	3.04	24,753	17.1	11	1
39 Pennsylvania	PA	12,709,630	0.95	27.1	40.2	3.1	26,374	10.2	9.3	1
40 Rhode Island	RI	1,052,886	0.93	30.3	39.6	3.19	27,667	12.2	**9.2**	1
41 South Carolina	SC	4,636,312	0.95	24.6	37.8	3.12	22,128	17.5	13.8	1
42 South Dakota	SD	816,463	1.02	26.3	37.3	3.07	23,647	12.4	9.2	3
43 Tennessee	TN	6,356,897	0.95	23.1	38	3.1	22,463	14.4	13.4	1
44 Texas	TX	25,257,114	0.98	25.9	33.6	3.41	23,863	23.7	13.8	1
45 Utah	UT	2,776,469	1.01	29.3	29.2	3.57	22,059	15.3	9.7	3
46 Vermont	VT	625,960	0.97	33.6	41.5	2.9	**26,876**	8	8.4	1
47 Virginia	VA	8,024,617	0.96	34.3	37.5	3.15	31,313	13.1	7.7	1
48 Washington	WA	6,744,496	0.99	31	37.2	3.11	28,364	14.2	9.2	1
49 West Virginia	WV	1,853,973	0.97	17.5	41.2	3	**20,953**	14.6	13.2	1
50 Wisconsin	WI	5,691,047	0.99	26.4	38.5	3.02	25,458	9.4	9.1	1
51 Wyoming	WY	564,460	1.04	24.1	36.7	3.02	27,616	14.9	7.2	3

Note. Abbr. = state name abbreviation. Population = state population. MaleRatio = ratio of number of males to number of females. Edu. = percentage of state population with four year college degree or higher among 25 years old and older individuals. Age = median age for state population. Income = inflation adjusted per capita income in 2010 dollars. NoIns. = percentage of state population without health insurance. PovRt = average state poverty rate. Grp. = grouping based on whether MaleRatio is less than, equal to, or larger than 1. The data shown in bold was set as missing for imputation purposes.

Mean Imputation

For mean imputation, we need to first obtain the means of all variables with missing data. In our example, the data is missing for Income and PovertyRate. The SPSS syntax and the corresponding default output needed to generate summary statistics for Income and PovertyRate are shown in Figure 31. From this output we see that these means are 26,205.09 and 10.53 respectively. The next step is to open the SPSS syntax

file through the "File > New > Syntax" menu command as indicated in Figure 32. Once the new syntax window opens up, we need to write four lines of code as shown in Figure 33.



	Education	Age	FamilySize	Income	NoInsurance	PovertyRate	Group
1	21.9	37.8	3.16	21993	14.6	.	1
2	27.9	33.8	3.24	30598	19.9	7.2	3
3	25.9	35.9	3.28	23618	16.9	12.5	1
4	19.5	37.3	3.09	20725	17.5	14.1	1
5	30.1	35.2	3.53	27353	18.5	11.8	1
6	36.4	36.0	3.12	28723	15.9	9.4	2
7	35.5	40.0	3.15	35078	9.1	.	1
8	27.8	38.8	3.24	27729	9.7	8.1	1
9	50.1	33.9	3.37	41240	7.6	14.1	1
10	25.8	40.7	3.24	24272	21.3	12.0	1
11	27.3	35.4	3.30	23383	19.7	13.7	1
12	29.5	38.7	3.59	27537	7.9	7.4	3
13	24.4	34.7	3.20	20991	17.7	11.6	2
14	30.7	36.6	3.28	27325	13.8	10.1	1
15	22.7	36.9	3.11	22806	14.8	11.0	1
16	25.0	38.2	2.97	.	9.3	8.2	1
17	29.8	36.2	3.12	24911	13.9	9.5	1
18	20.5	38.0	3.06	21706	15.3	14.5	1
19	21.4	35.9	3.21	22862	17.8	14.5	1
20	26.8	42.7	2.90	24950	10.1	8.8	1

Figure 30. Data from Table 29 with missing values in SPSS data editor.

After writing the code, next task is to run the it using the "Run > All" menu command in the syntax window as shown in Figure 34. This action automatically replaces the missing data for each variable with the corresponding mean values. A screenshot of complete data is shown in Figure 35. A comparison between Figures 30 and 35 shows that missing data has been replaced by the mean values for both Income and PovertyRate. It should be noted that the dataset with imputed values should be saved with a different name in order to keep the dataset containing missing values available for remaining imputation methods.

```
DESCRIPTIVES VARIABLES=Income PovertyRate
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

```
[DataSet1] D:\02 - Data - With Missing.sav
```

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Inflation adjusted per capita income in 2010 dollars	46	19096	41240	26205.09	4405.180
Average poverty rate	46	5.3	17.8	10.53	2.6947
Valid N (listwise)	41				

Figure 31. SPSS syntax needed to compute means of Income and PovertyRate with the corresponding output.

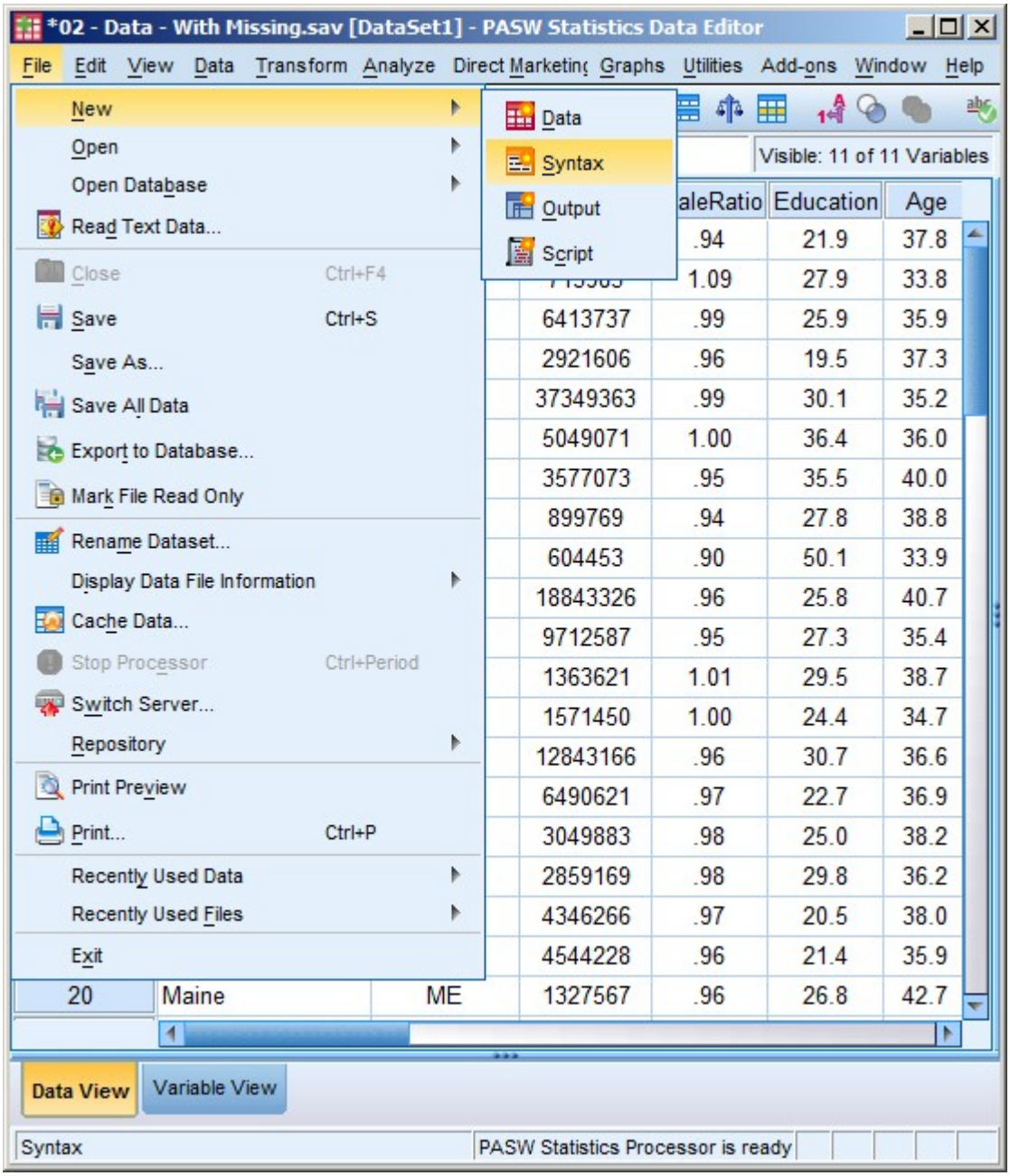


Figure 32. Steps required to open a new syntax file in SPSS.

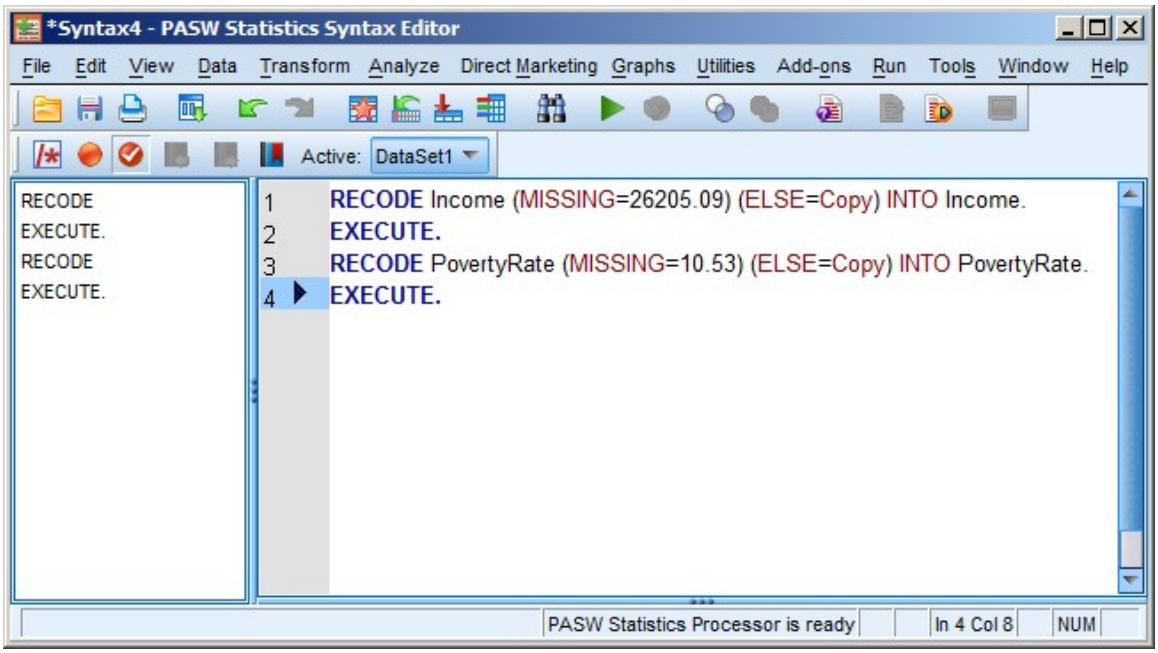


Figure 33. SPSS code needed to impute missing data for variables Income and PovertyRate.

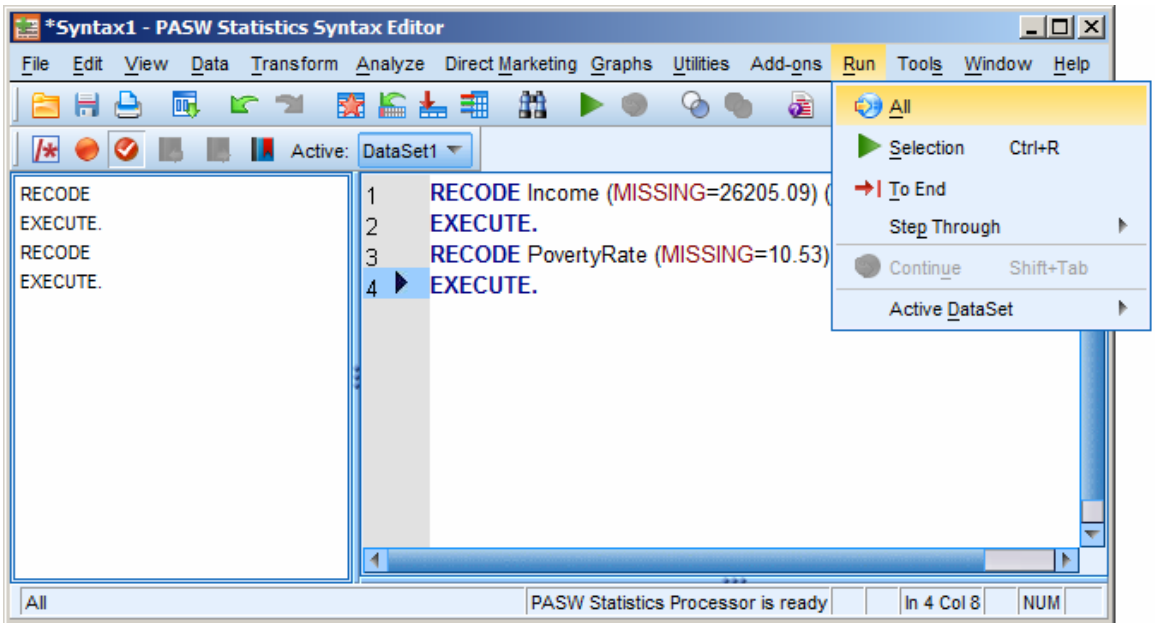
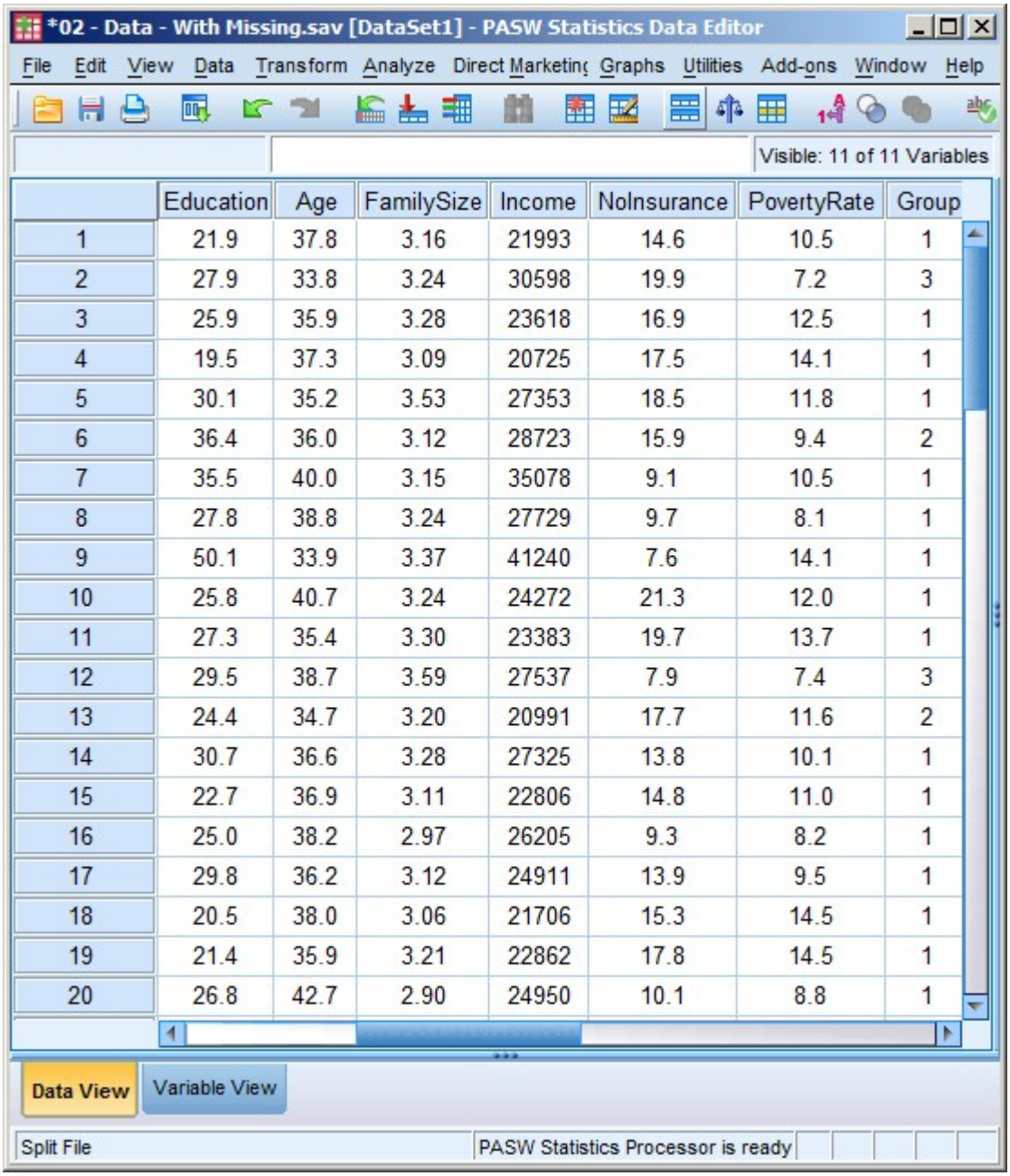


Figure 34. Steps required to run SPSS code.



	Education	Age	FamilySize	Income	NoInsurance	PovertyRate	Group
1	21.9	37.8	3.16	21993	14.6	10.5	1
2	27.9	33.8	3.24	30598	19.9	7.2	3
3	25.9	35.9	3.28	23618	16.9	12.5	1
4	19.5	37.3	3.09	20725	17.5	14.1	1
5	30.1	35.2	3.53	27353	18.5	11.8	1
6	36.4	36.0	3.12	28723	15.9	9.4	2
7	35.5	40.0	3.15	35078	9.1	10.5	1
8	27.8	38.8	3.24	27729	9.7	8.1	1
9	50.1	33.9	3.37	41240	7.6	14.1	1
10	25.8	40.7	3.24	24272	21.3	12.0	1
11	27.3	35.4	3.30	23383	19.7	13.7	1
12	29.5	38.7	3.59	27537	7.9	7.4	3
13	24.4	34.7	3.20	20991	17.7	11.6	2
14	30.7	36.6	3.28	27325	13.8	10.1	1
15	22.7	36.9	3.11	22806	14.8	11.0	1
16	25.0	38.2	2.97	26205	9.3	8.2	1
17	29.8	36.2	3.12	24911	13.9	9.5	1
18	20.5	38.0	3.06	21706	15.3	14.5	1
19	21.4	35.9	3.21	22862	17.8	14.5	1
20	26.8	42.7	2.90	24950	10.1	8.8	1

Figure 35. SPSS data editor showing complete data after mean imputation.

Regression imputation

For regression imputation, open the dataset that has missing data. Then use the "Analyze > Missing Value Analysis" option from menu bar (see Figure 36) to open the "Missing Value Analysis" window (see Figure 37). The next step is to specify the variables that SPSS should use for missing data imputation. This requires specifying not only those variables that have missing data but also any other variables that are available in the dataset and that are related to the variables with missing data. In our example we chose to use all of the available quantitative variables for this purpose. All scale variables should be moved to the Quantitative Variables box and all nominal or ordinal variables to the Categorical Variables box as shown in Figure 38. Notice that after this step is completed, the check boxes next to the various estimation boxes become available. From these, select Regression as shown in Figure 39 and then click the Variables button. This opens the "Missing Value Analysis: Regression" window as shown in Figure 40. In this window, check the "Save completed data" check box (see Figure 41) and then type a name for the complete dataset, including imputed values that will be generated by SPSS. We have used the name Regression_Imputed_Dataset for this complete dataset in this example. Next click Continue to go back to the Missing value Analysis box. Click OK to continue. This will result in a new data editor window opening in SPSS that contains the completed dataset (see Figure 42). This new dataset can now be saved separately from the original data file.

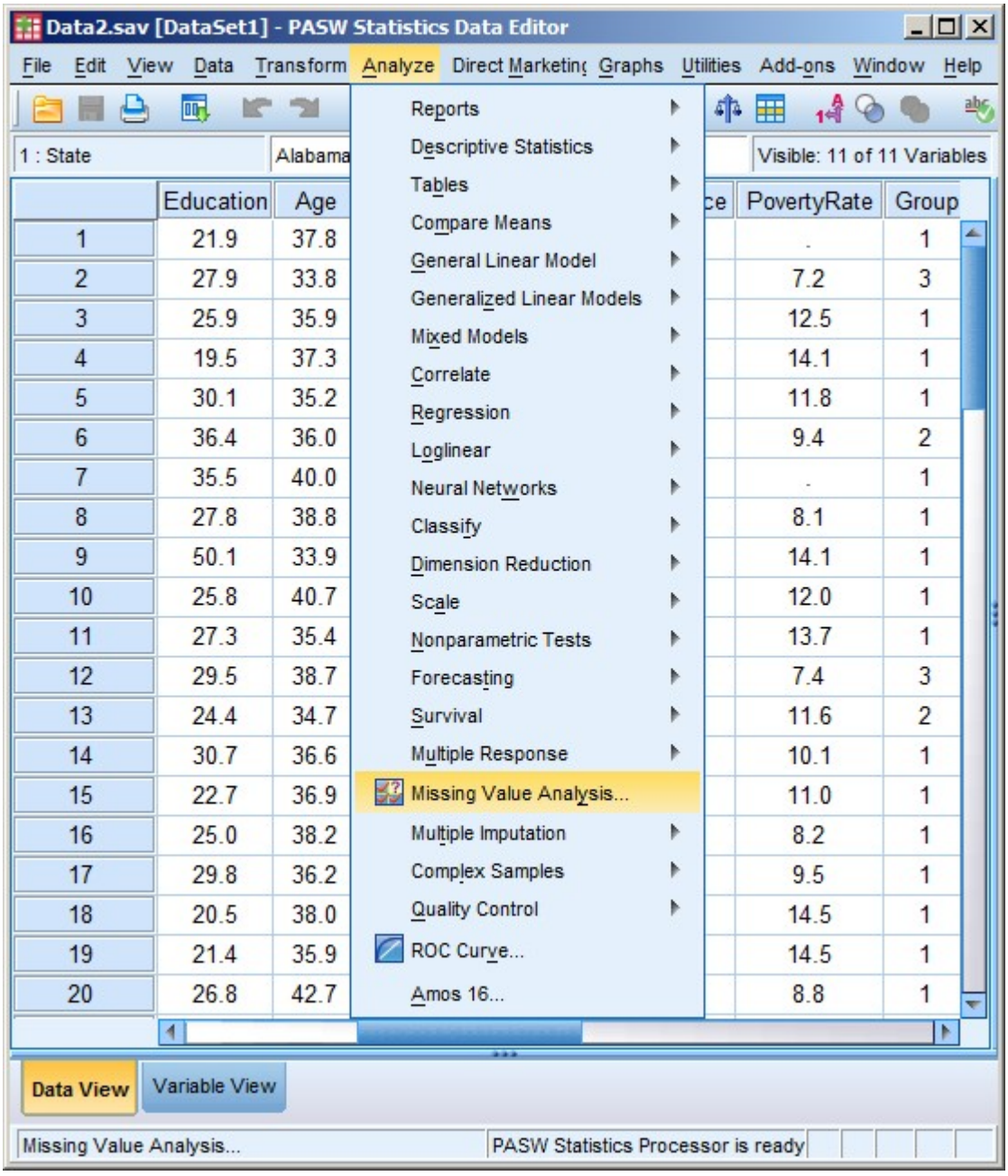


Figure 36. The Missing Value Analysis menu option.

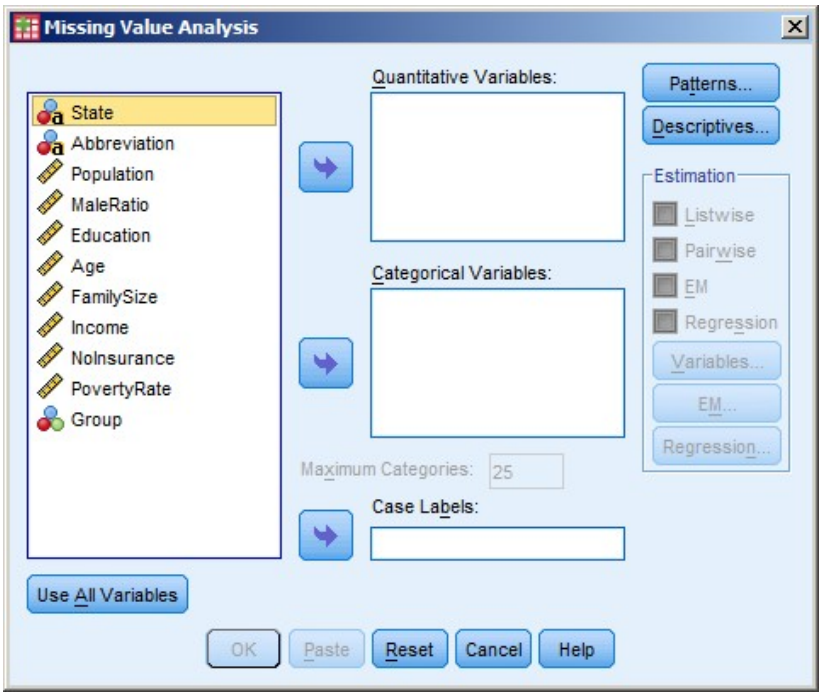


Figure 37. Missing Value Analysis window with initial position of variables.

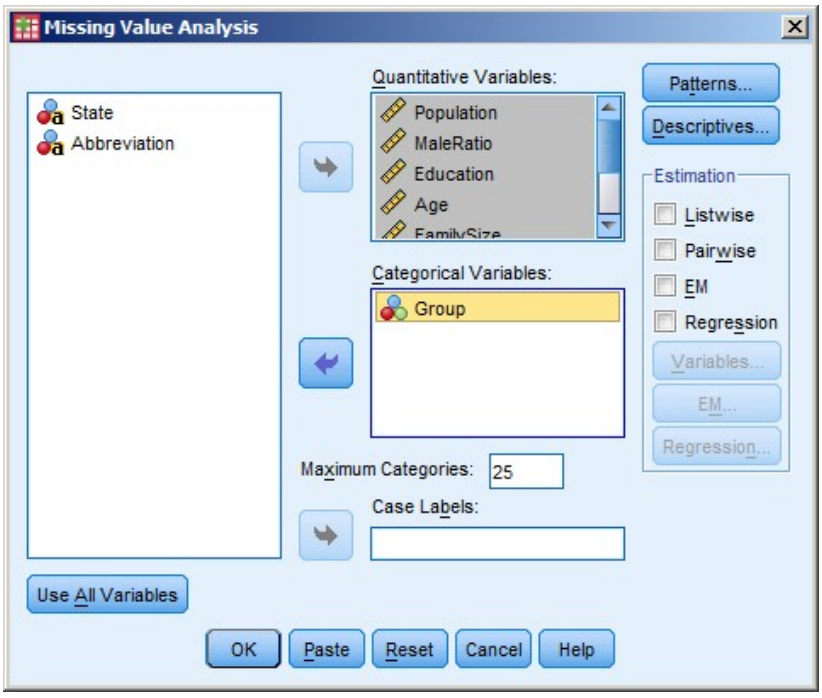


Figure 38. Missing Value Analysis Window with final position of variables.

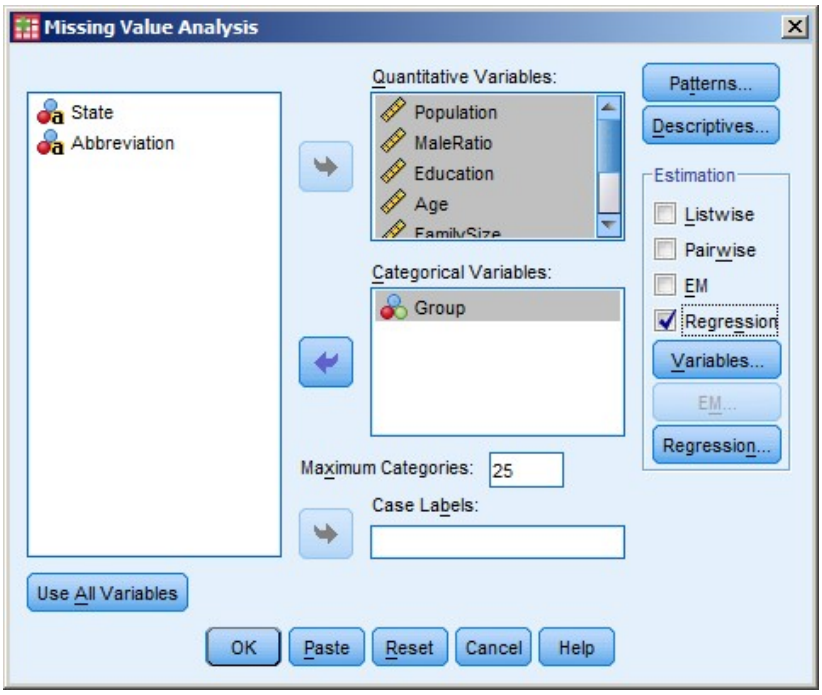


Figure 39. Missing Value Analysis window with regression option checked.

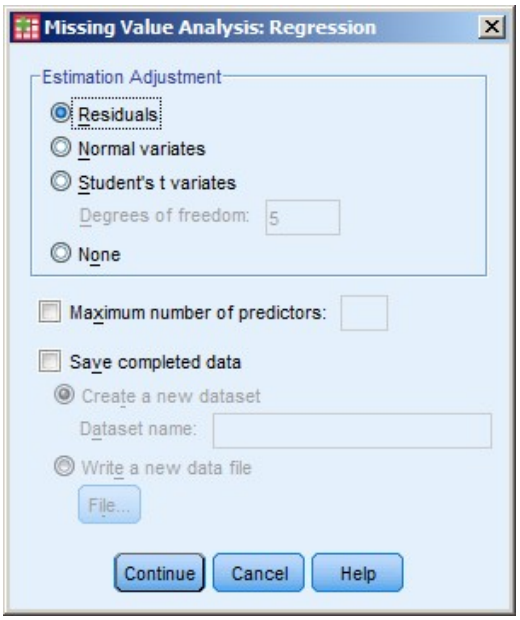


Figure 40. Missing Value Analysis: Regression window with default options.

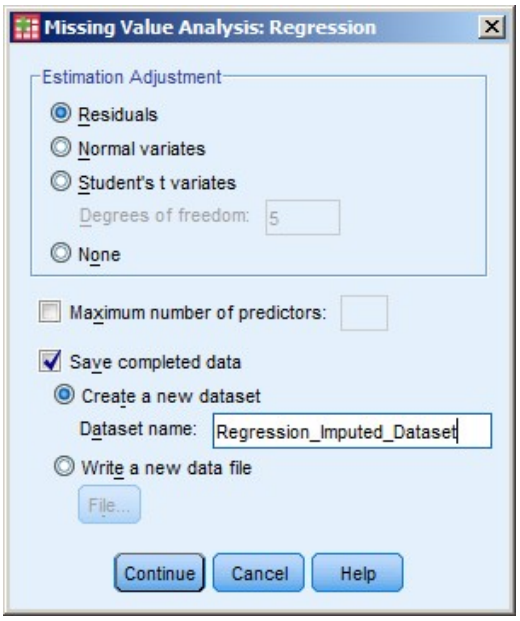
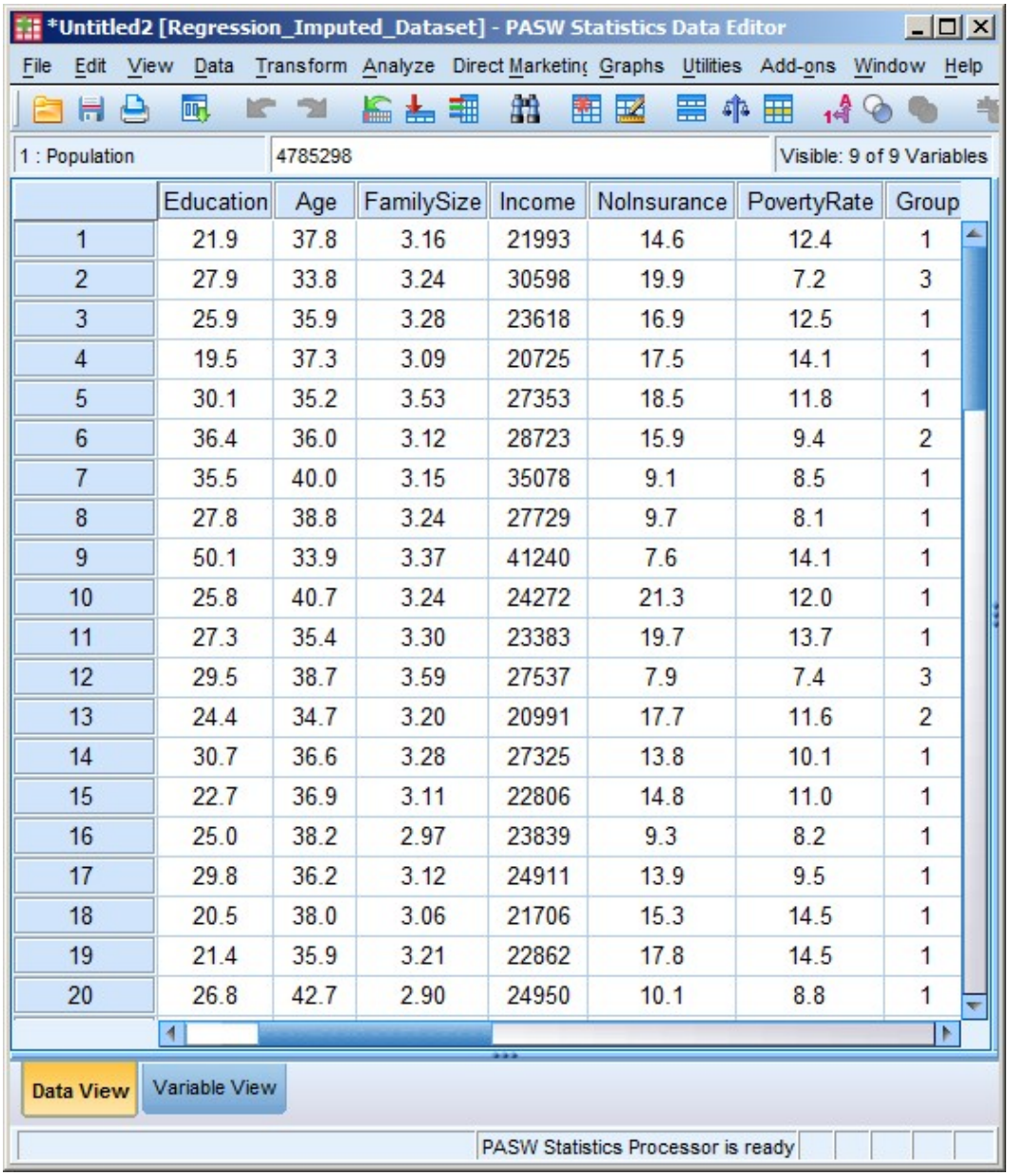


Figure 41. Missing Value Analysis: Regression window with selected options.

EM Imputation

The steps required for EM imputation in SPSS are similar to those outlines earlier for regression imputation. The starting point is the dataset with missing data. Once the data is open in the data editor, open the Missing Value Analysis box and specify the continuous and categorical variable to be used in missing data imputation (Figures 36, 37, and 38). In the estimation options, check the check box next to EM (Figure 43), then click the EM button. This will open the Missing Value Analysis: EM window (Figure 44). In the Missing Value Analysis: EM window, check the check box next to "Save completed data" and select a name for the completed dataset that SPSS will generate (Figure 45). We have used the name EM_Imputed_Dataset in this example. Next click

Continue to go back to Missing Value Analysis window. Hitting OK in this window will generate the completed dataset by opening a new data editor window (Figure 46).



	Education	Age	FamilySize	Income	NoInsurance	PovertyRate	Group
1	21.9	37.8	3.16	21993	14.6	12.4	1
2	27.9	33.8	3.24	30598	19.9	7.2	3
3	25.9	35.9	3.28	23618	16.9	12.5	1
4	19.5	37.3	3.09	20725	17.5	14.1	1
5	30.1	35.2	3.53	27353	18.5	11.8	1
6	36.4	36.0	3.12	28723	15.9	9.4	2
7	35.5	40.0	3.15	35078	9.1	8.5	1
8	27.8	38.8	3.24	27729	9.7	8.1	1
9	50.1	33.9	3.37	41240	7.6	14.1	1
10	25.8	40.7	3.24	24272	21.3	12.0	1
11	27.3	35.4	3.30	23383	19.7	13.7	1
12	29.5	38.7	3.59	27537	7.9	7.4	3
13	24.4	34.7	3.20	20991	17.7	11.6	2
14	30.7	36.6	3.28	27325	13.8	10.1	1
15	22.7	36.9	3.11	22806	14.8	11.0	1
16	25.0	38.2	2.97	23839	9.3	8.2	1
17	29.8	36.2	3.12	24911	13.9	9.5	1
18	20.5	38.0	3.06	21706	15.3	14.5	1
19	21.4	35.9	3.21	22862	17.8	14.5	1
20	26.8	42.7	2.90	24950	10.1	8.8	1

Figure 42. Complete dataset with regression imputed values.

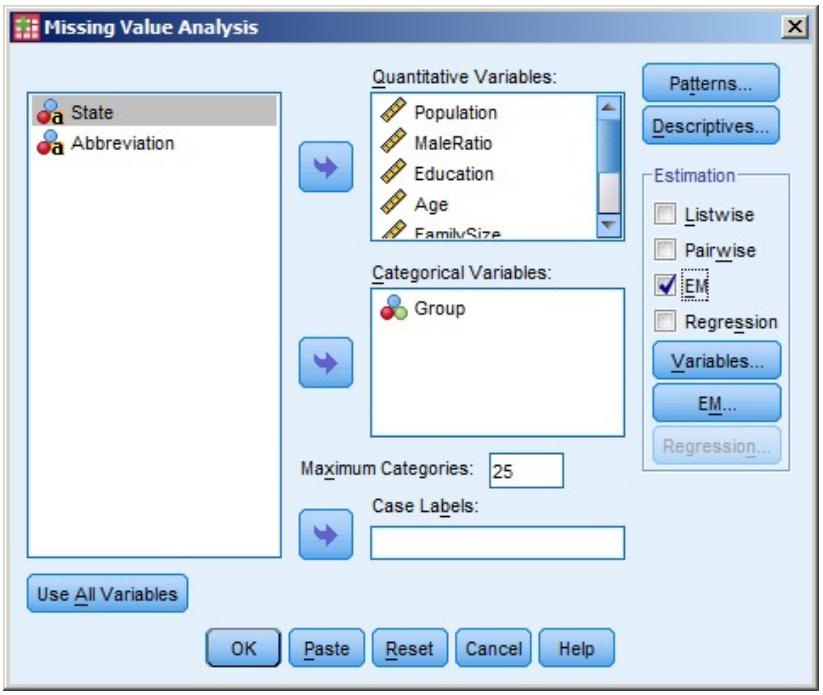


Figure 43. Missing Value Analysis window with EM option checked.

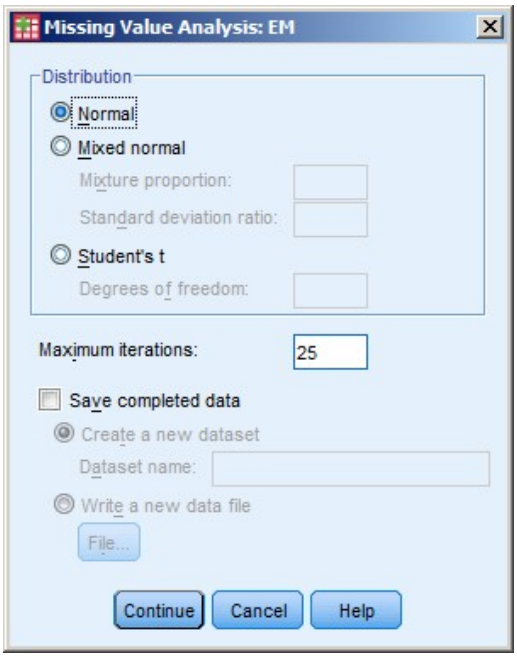


Figure 44. Missing Value Analysis: EM window with default options.

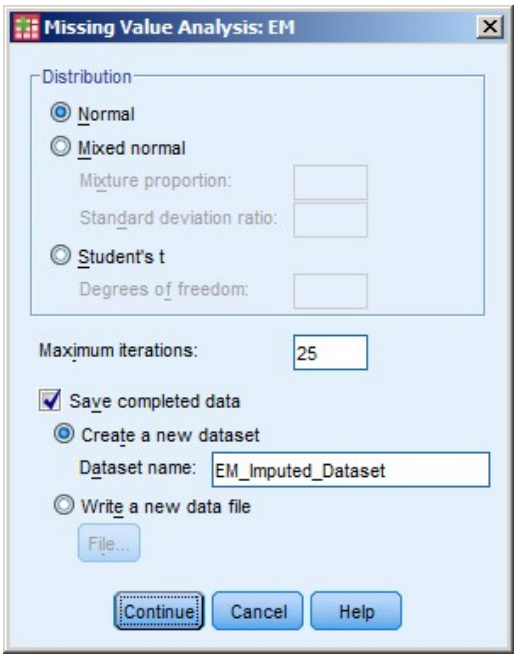
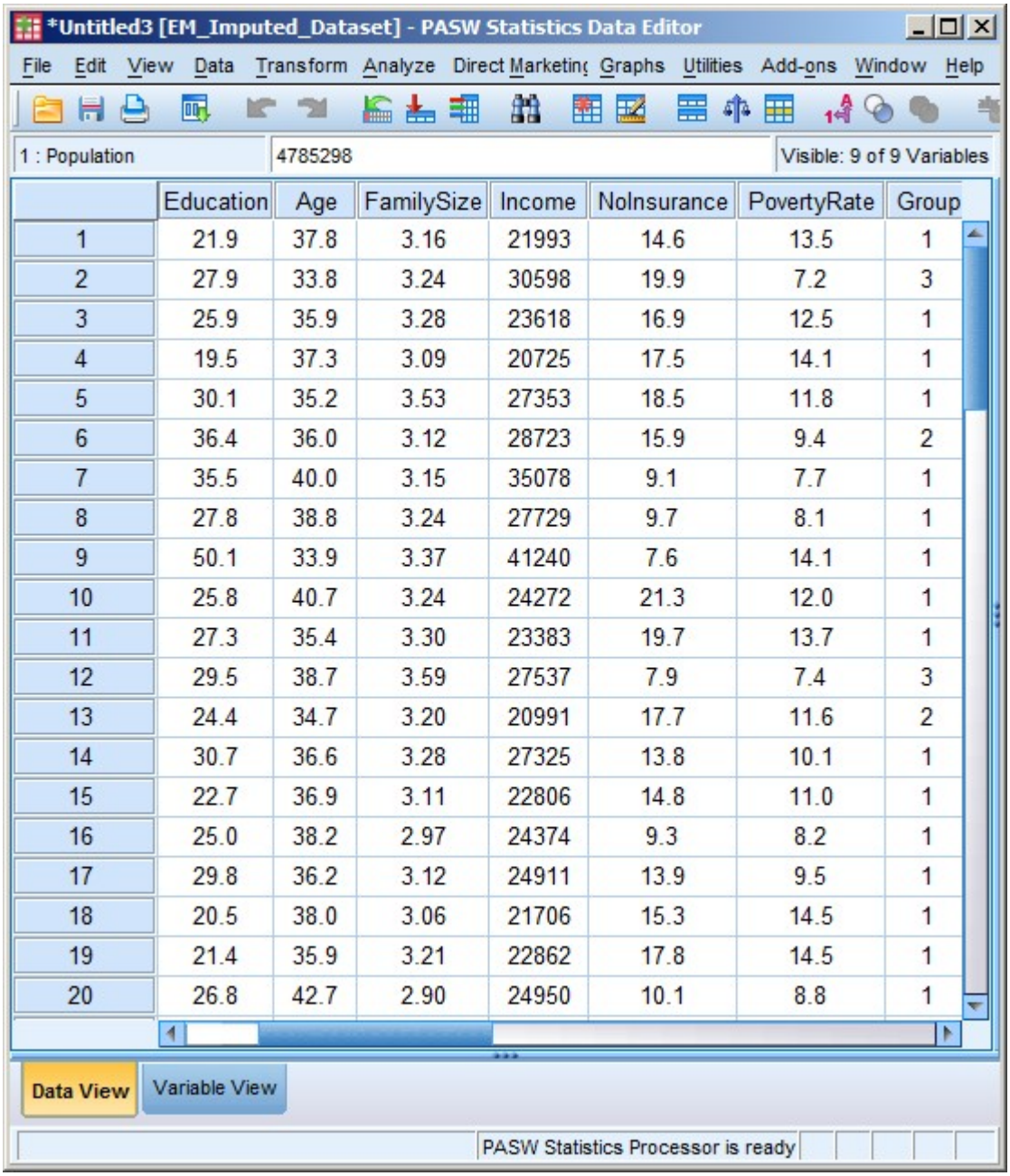


Figure 45. Missing Value Analysis: EM window with selected options.

Multiple Imputation

The multiple imputation command can be accessed through the "Analyze > Multiple Imputation > Impute Missing Data Values" menu option in SPSS (Figure 47). This action will open the Impute Missing Data Values window (Figure 48). Next, move all variables (continuous and categorical) to the Variables in Model box and provide a name for the completed dataset that SPSS will generate (see Figure 49). In this example we have given the name Multiple_Imputed_Dataset to our completed dataset. We will keep the default number of imputations, 5, unchanged. Clicking OK now at the bottom of this window will open a new data editor window containing original and five versions of completed dataset, one for each imputation (Figure 50).



	Education	Age	FamilySize	Income	NoInsurance	PovertyRate	Group
1	21.9	37.8	3.16	21993	14.6	13.5	1
2	27.9	33.8	3.24	30598	19.9	7.2	3
3	25.9	35.9	3.28	23618	16.9	12.5	1
4	19.5	37.3	3.09	20725	17.5	14.1	1
5	30.1	35.2	3.53	27353	18.5	11.8	1
6	36.4	36.0	3.12	28723	15.9	9.4	2
7	35.5	40.0	3.15	35078	9.1	7.7	1
8	27.8	38.8	3.24	27729	9.7	8.1	1
9	50.1	33.9	3.37	41240	7.6	14.1	1
10	25.8	40.7	3.24	24272	21.3	12.0	1
11	27.3	35.4	3.30	23383	19.7	13.7	1
12	29.5	38.7	3.59	27537	7.9	7.4	3
13	24.4	34.7	3.20	20991	17.7	11.6	2
14	30.7	36.6	3.28	27325	13.8	10.1	1
15	22.7	36.9	3.11	22806	14.8	11.0	1
16	25.0	38.2	2.97	24374	9.3	8.2	1
17	29.8	36.2	3.12	24911	13.9	9.5	1
18	20.5	38.0	3.06	21706	15.3	14.5	1
19	21.4	35.9	3.21	22862	17.8	14.5	1
20	26.8	42.7	2.90	24950	10.1	8.8	1

Figure 46. Complete dataset with EM imputed values.

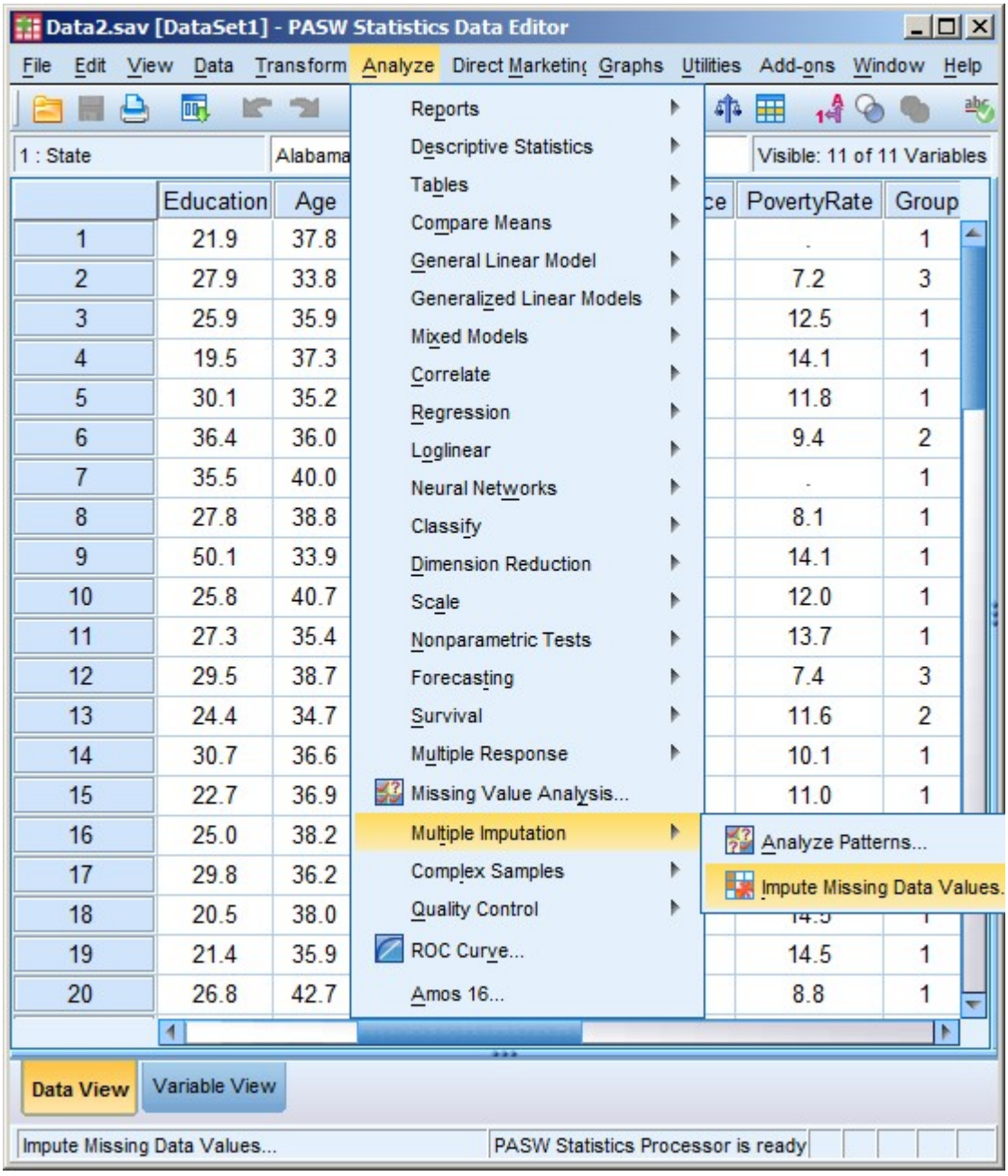


Figure 47. Multiple Imputation menu option in SPSS.

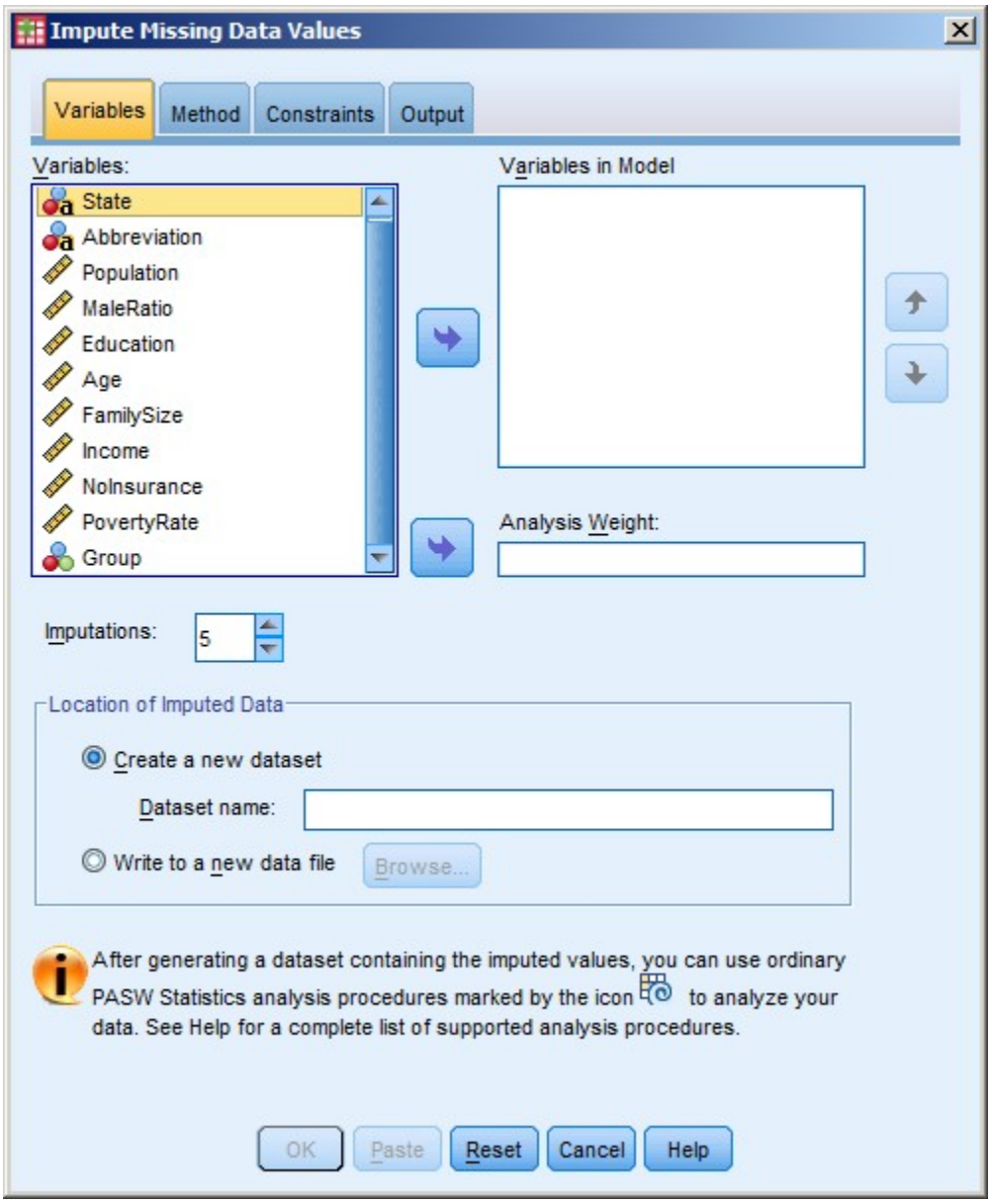


Figure 48. Impute Missing Data Values window, initial view.

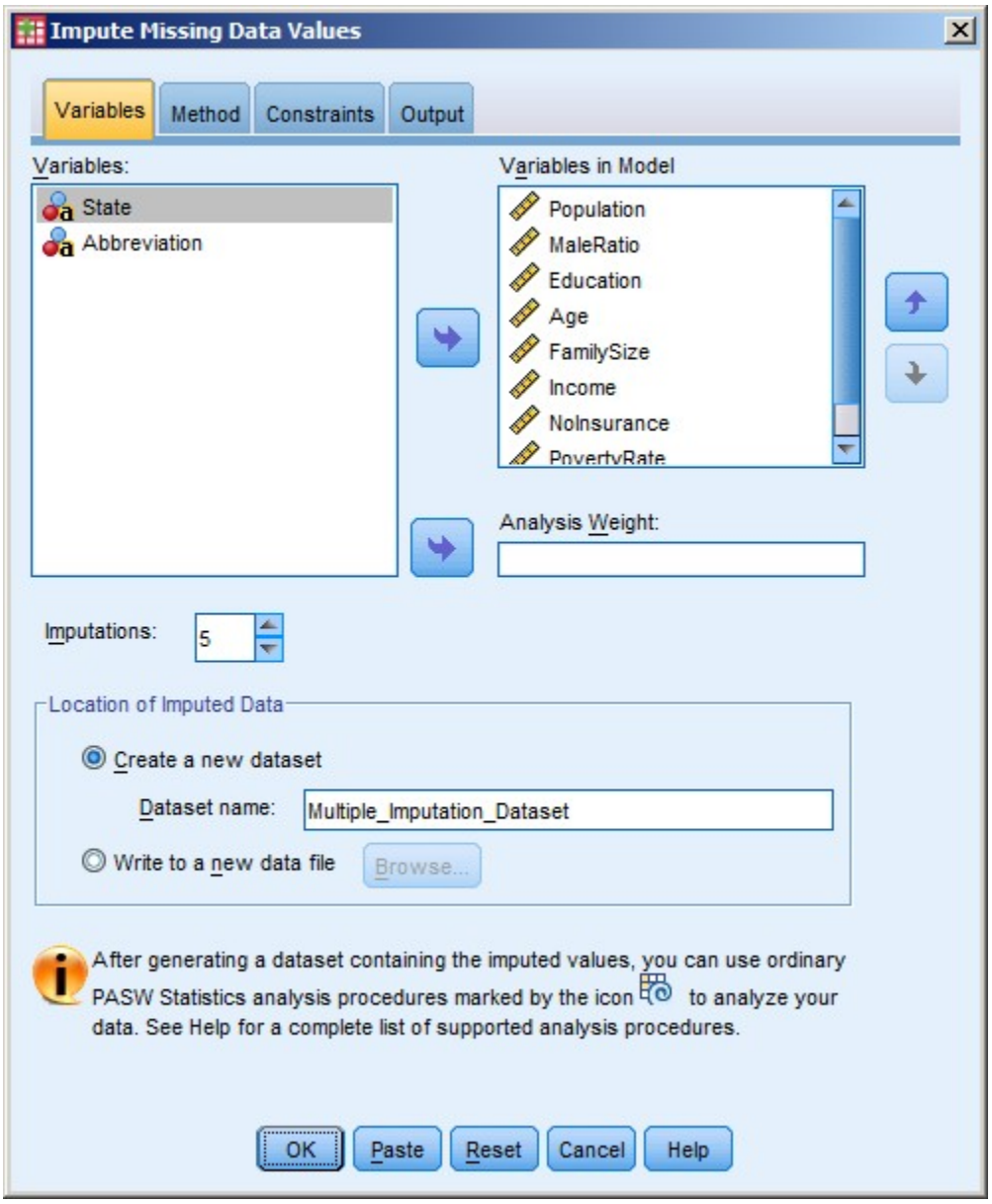


Figure 49. Impute Missing Data Values window, final view.

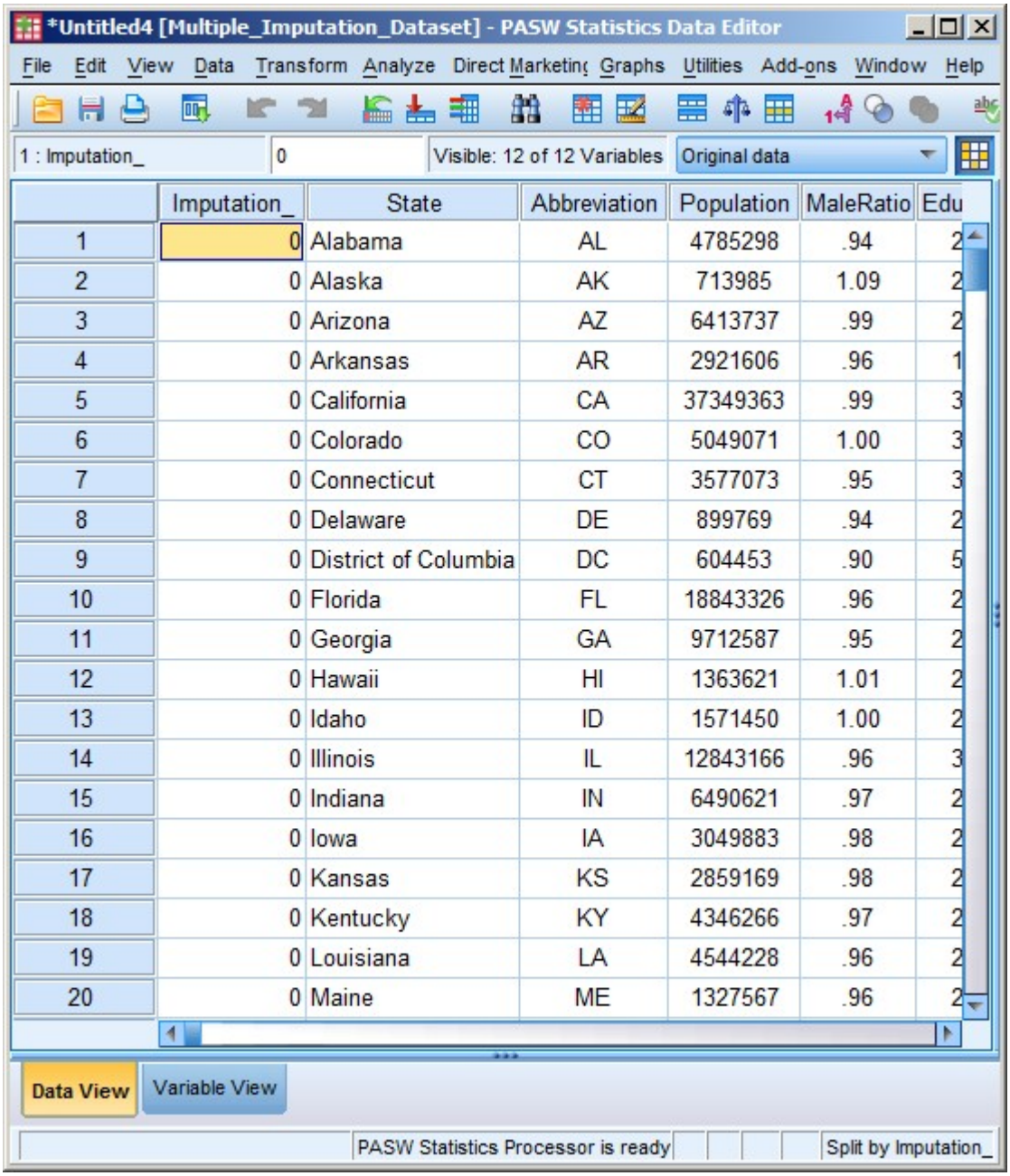


Figure 50. SPSS data editor showing completed multiple imputation dataset

It is important to note here that although we have generated the completed datasets under multiple imputation, the dataset is not yet ready for analysis. The imputed datasets are stacked on top of each other. This can be seen clearly by scrolling to the bottom of data editor window containing imputed data (see Figure 51). The total number of rows shows is 306 i.e. 51 original rows of data plus 51 x 5 = 255 rows generated by the five completed datasets. In order to perform statistical analyses separately on these six datasets, we need to inform SPSS that they form different groups of observations. This can be done through the "Data > Split File" menu option (see Figure 52). In the Split File window, choose Compare Groups radio button and move Imputation_ from variable list to "Groups based on" box (Figures 53 and 54). Next, click OK. From now on any analysis performed in SPSS will be repeated for each group with results for all groups displayed side by side. For illustration we ran the descriptive statistics procedure on Income. Results are provided in Figure 55. It should be noted that although parameter estimates are provided for each completed dataset by SPSS, it remains to the analyst to manually combine those statistics using appropriate formulas.

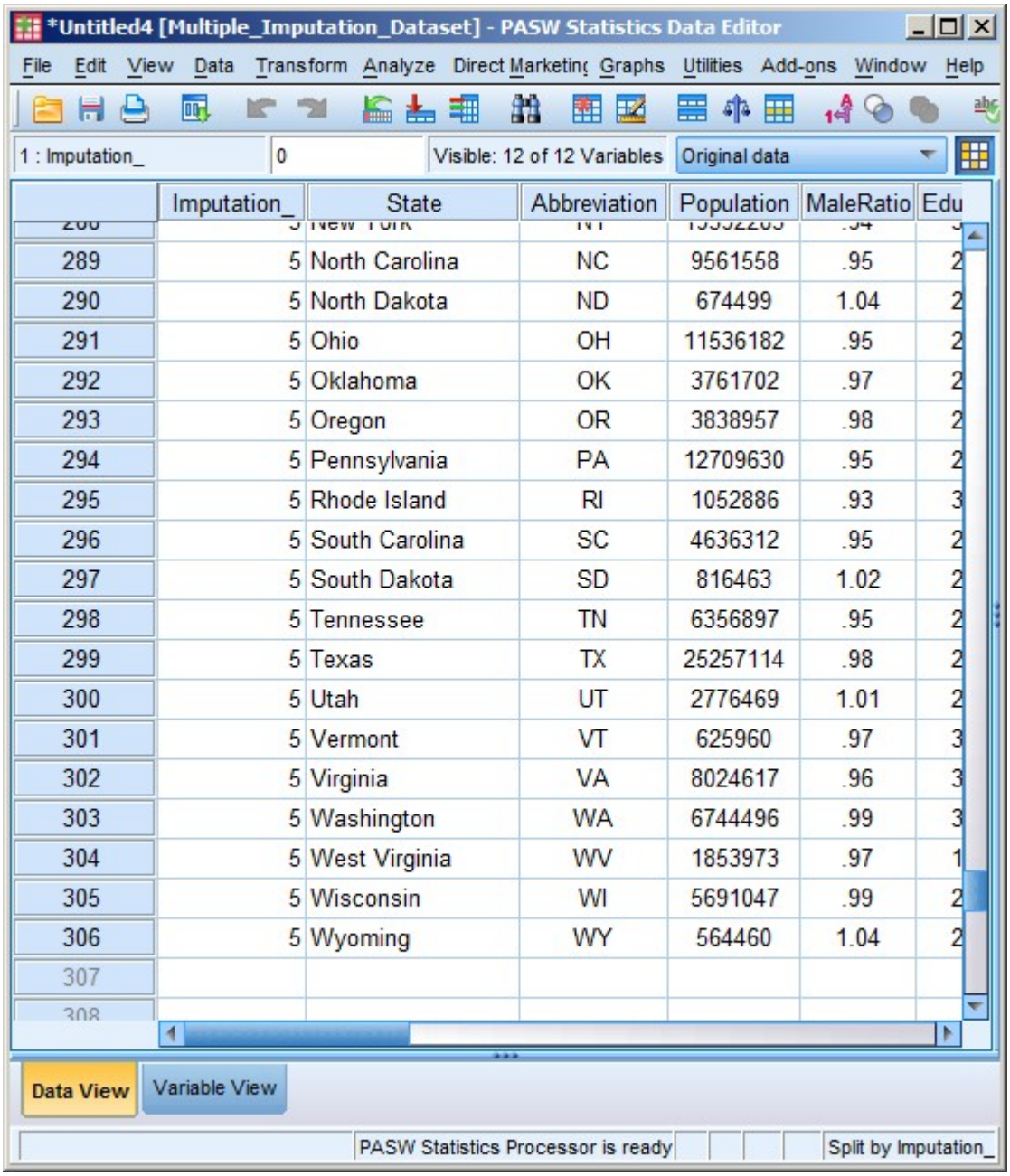


Figure 51. Data editor window in SPSS showing how completed datasets after multiple imputation are stacked on top of each other.

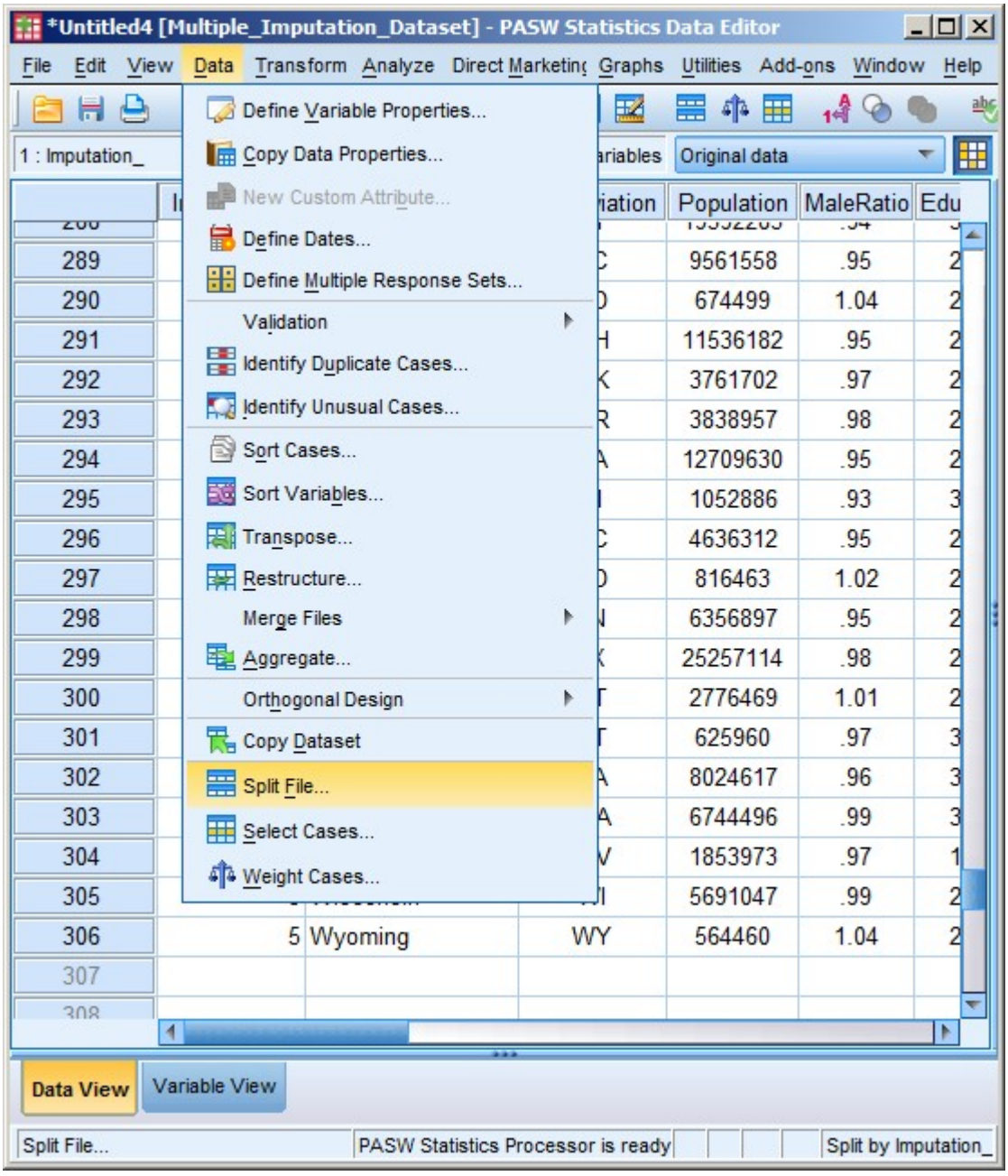


Figure 52. Split File menu option in SPSS.

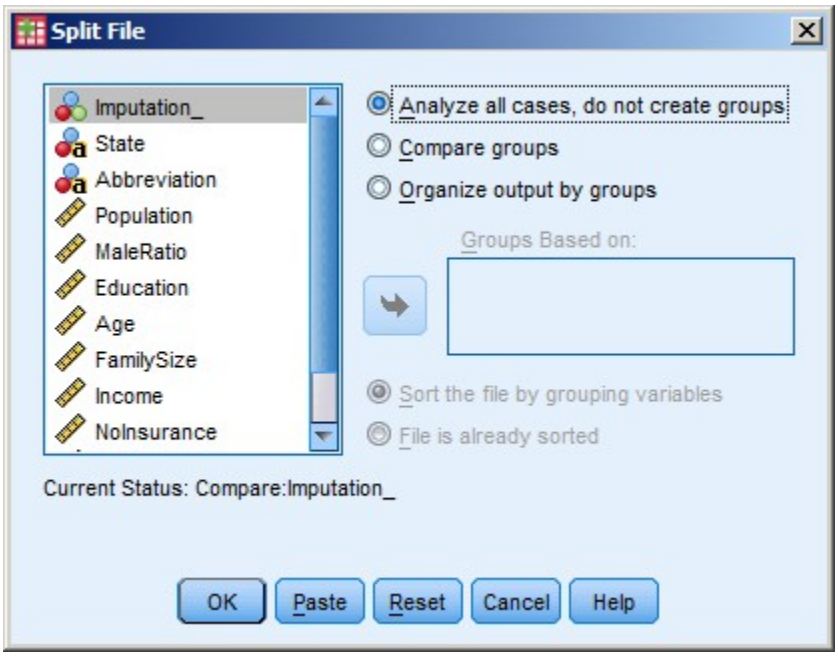


Figure 53. Split File window, initial view.

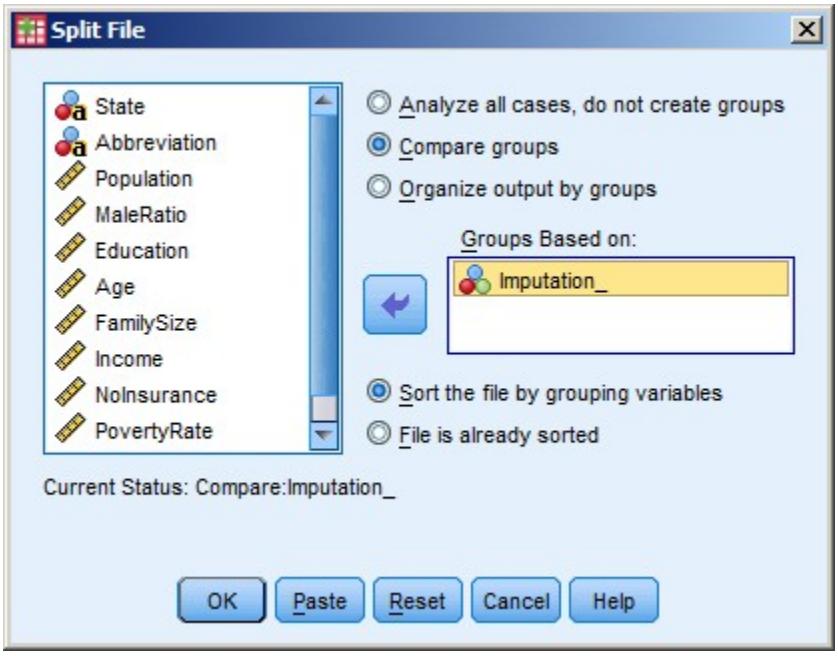


Figure 54. Split File window, final view.

```
DESCRIPTIVES VARIABLES=Income
  /STATISTICS=MEAN STDDEV MIN MAX.
```

Descriptives

```
[Multiple_Imputation_Dataset]
```

Descriptive Statistics

Imputation Number		N	Minimum	Maximum	Mean	Std. Deviation
Original data	Inflation adjusted per capita income in 2010 dollars	46	19096	41240	26205.09	4405.180
	Valid N (listwise)	46				
1	Inflation adjusted per capita income in 2010 dollars	51	18116	41240	25782.94	4605.006
	Valid N (listwise)	51				
2	Inflation adjusted per capita income in 2010 dollars	51	19096	41240	26153.28	4289.771
	Valid N (listwise)	51				
3	Inflation adjusted per capita income in 2010 dollars	51	17244	41240	26004.97	4545.826
	Valid N (listwise)	51				
4	Inflation adjusted per capita income in 2010 dollars	51	15955	41240	25921.30	4477.762
	Valid N (listwise)	51				
5	Inflation adjusted per capita income in 2010 dollars	51	18086	41240	25861.54	4415.702
	Valid N (listwise)	51				

Figure 55. SPSS syntax needed to compute descriptive statistics for variable Income. Notice that the analysis is conducted separately by SPSS for original and imputed datasets.

References

References

Acock, A. (2005). SAS, Stata, SPSS: A comparison. *Journal of Marriage and Family, 67*(4), 1093-1095. doi:10.1111/j.1741-3737.2005.00196.x

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

Agresti, A. (2007). *An introduction to categorical data analysis*. Hoboken, NJ: John Wiley & Sons, Inc. doi:10.1002/0470114754

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

Baker, F. (2001). *The basics of item response theory*. ERIC Clearinghouse on Assessment and Evaluation.

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

Buck, S. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society. Series B, Methodological, 22*(2), 302-306.

Chen, S., Jain, L., & Tai, C. (2006). *Computational economics: A perspective from computational intelligence*. Hershey, PA: Idea Group, Inc. doi: 10.4018/978-1-59140-649-5

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B, (Methodological) 39*(1), 1-38.

Edgett, G. (1956). Multiple regression with missing observations among the independent variables. *Journal of the American Statistical Association, 51*(273), 122-131. doi:10.2307/2280808

Embretson, S., & Reise S. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

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

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

Gujarati, D. (2003). *Basic econometrics*. New York, NY: McGraw-Hill/Irwin.

Gurland, J., & Tripathi, R. (1971). A simple approximation for unbiased estimation of the standard deviation. *The American Statistician, 25*(4), 30-32. doi:10.2307/2682923

Hahs-Vaughn, D. (2005). A primer for using and understanding weights with national datasets. *The Journal of Experimental Education, 73*(3), 221-248. doi:10.3200/JEXE.73.3.221-248

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

Hambleton, R., Swaminathan, H. (1985). *Item response theory: Principles and applications.* Norwell, MA: Kluwer-Nijhoff Publishing.

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

Hoenig, J., & Heisey, D. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician, 55*(1), 19-24. doi:10.1198/000313001300339897

Hong, T., & Wu, C. (2011). Mining rules from an incomplete dataset with a high missing rate. *Expert Systems with Applications, 38*, 3931-3936. doi:10.1016/j.eswa.2010.09.054

Howell, D. (2007). *Statistical methods for psychology*. Belmont, CA: Thompson Wadsworth.

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

Kennedy, C. (2005). *Single-case designs for educational research.* Boston, MA: Allyn & Bacon.

Kim, J. (2004). Finite sample properties of multiple imputation estimators. *The Annals of Statistics, 32*(2), 766-783. doi:10.1214/009053604000000175

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

Lord, F. (1980). *Applications of item response theory to practical testing problems.* New Jersey: Lawrence Erlbaum Associates, Inc.

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.

McKnight, P., McKnight, K., Sidani, S., & Figueredo, A. (2007). *Missing data: A gentle introduction.* New York, NY: The Guilford Press.

Mislevy, R. (1991). Randomization-based inference about latent variable from complex samples. *Psychometrika, 56*(2), 177–196. doi:10.1007/BF02294457

Mislevy, R., Beaton, A., Kaplan, B., & Sheehan, K. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*(2), 133–161. doi:10.1111/j.1745-3984.1992.tb00371.x

Murtonen, M., & Lethtinen, E. (2003). Difficulties experienced by education and sociology students in quantitative methods courses. *Studies in Higher Education, 28*(2), 171-185. doi:10.1080/0307507032000058064

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

Organization for Economic Cooperation and Development (OECD) (2005). *PISA 2003: Technical Report.* Paris: OECD Publishing. doi:10.1787/9789264010543-en

Park, H. (2008). *Hypothesis Testing and Statistical Power of a Test.* Working Paper. The University Information Technology Services (UITS) Center for Statistical and Mathematical Computing, Indiana University. Retrieved from http://www.indiana.edu/~statmath/stat/all/power/power.pdf

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. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. doi:10.2307/2335739

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

Salkind, N., & Rasmussen, K. (2007). *Encyclopedia of measurement and statistics.* Thousand Oaks, CA: Sage Publications, Inc.

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

Taguchi, K., Funama, Y., Zhang, M., Fishman, E., & Geschwind J. (2009). Quantitative measurement of iodine concentration in the liver using abdominal C-arm computed tomography. *Academic Radiology, 16*(2), 200-208. doi:10.1016/j.acra.2008.08.002

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. (2001). Missing value estimation methods for DNA microarrays. *Bioinformatics, 17*(6), 520-525. doi:10.1093/bioinformatics/17.6.520a

UCLA. (2010). *Resources to help you learn and use SPSS*. UCLA: Academic Technology Services, Statistical Consulting Group. Retrieved from http://www.ats.ucla.edu/stat/spss

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.

Weisstein, E. (1999). *CRC Concise Encyclopedia of mathematics*. Boca Raton, FL: CRC Press LLC. doi:10.1201/9781420035223

Wilks, S. (1932). Moments and distributions of estimates of population parameters from fragmentary samples. *The Annals of Mathematical Statistics, 3*(3), 163-195. doi:10.1214/aoms/1177732885

Wu, M. (2005). The role of plausible values in large-scale surveys. *Studies in Educational Evaluation, 31*(2), 114–128. doi:10.1016/j.stueduc.2005.05.005

Yesilova, A., Kaya, Y., & Almali, M. (2011). A comparison of hot deck imputation and substitution methods in the estimation of missing data. *Gazi University Journal of Science, 24*(1), 69-75.

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

Zhou, X., Wang, X., & Dogherty, E. (2003). Missing-value estimation using linear and non-linear regression with Bayesian gene selection. *Bioinformatics, 19*(17), 2302-2307. doi:10.1093/bioinformatics/btg323

Curriculum Vitae

Jehanzeb Cheema is originally from Pakistan and earned his Doctorate in Economics in 2006, with specializations in Labor Economics and Development Economics, from University of Wisconsin-Milwaukee. He received his Doctorate in Education with a specialization in Research Methodology in 2012. Jehanzeb plans to stay in academia and teach quantitative courses in Education.