# Best Practice in Handling Cases of Missing or Incomplete Values in Data Analysis: A Guide against Eliminating Other Important Data

**Sub-theme: Improving Test Development Procedures to Improve Validity**

**'Dibu Ojerinde, Kunmi Popoola, Patrick Onyeneho, Chidinma Ifewulu**

**dibu65ojerinde@yahoo.com, kunmipopoola@yahoo.com, patrickonyeneho@yahoo.co.uk, chyfewulu@yahoo.com**

**Joint Admissions and Matriculation Board (JAMB), Abuja, Nigeria**

## Abstract

A common issue encountered in data analysis is the presence of missing values in datasets. Modern statistical techniques of data cleaning require complete data, but some statistical packages often default to the least desirable options in handling missing data such as (exclude cases list wise, exclude cases pair wise, exclude cases analysis by analysis, etc.,). Allowing software packages to perform the task of removing incomplete data most often creates the problem of eliminating a good deal of other important data that contribute to the overall analysis results. Incomplete or missing data affects the precision and validity of the result estimation depending on the extent of 'missingness'. Various methods are available for handling missing values before data analysis. This study is aimed at comparing the results of using complete data in analysis, data missing completely at random (MCAR) or missing at random (MAR), means substitution (MS), strong and weak imputations as well as multiple imputations (MI). With a random sample of 3,000 examinee responses in the UTME Physics that was extracted and analyzed, about 20% of the data were deleted to simulate an MCAR situation. When the Physics scores were correlated with the UTME aggregate scores, result obtained showed significant relationship as moderated by discipline applied by the examinees in the original dataset. Missing data corrections that use mean substitution, deletion method and weak or strong imputation methods were found to be biased with population parameters being overestimated or under estimated. MI yielded a closest significant estimate of the of the population parameter (at $p < 0.05$). It is recommended that MI method be used in handling missing or incomplete data because of the promise of providing unbiased estimates of relationship in MNAR situations. Missing values or incomplete data should not be discarded in analysis because doing this will limit the sample size, degree of freedom for analysis as well as produce biased estimates of the population parameters.

**Keywords: UTME, missingness, dataset, multiple imputations, unbiased estimate**

## Introduction

Incomplete or missing data is a common occurrence in many studies involving the use of quantitative data. Many datasets collected through survey and use of questionnaire often default to the least by having some variables with missing values from some research participants. When data on any variable from any participant is not available, that dataset is said to contain missing or incomplete data. While it is possible for some missing data to be so genuinely, there are situations in which missing data occur as a result of carelessness on the part of the researcher especially during data dressing or cleaning. Cole (2008) argued that in many studies carried out, that only a selected few authors seem to clearly explain how they handled the issue of missing data in their work, despite its obvious possibility to substantially skewing the results. It is therefore important in any study to

explain if missing data were ignored, substituted or that the researcher deleted cases with missing values list-wise, pair wise, or analysis by analysis as the case may be. These methods of handling missing data are often not the best practice and sometimes ineffective in handling incomplete data (Schafer & Graham, 2002).

It is often important to explain the rationale behind missing data in a dataset since there are various reasons for data to be missing as well as categories of missing data conditions. Data could be missing as a result of purposeful nonresponse or due to random influences. A female respondent may ignore answering certain questions because of its sensitivity, but the same is not true for a male respondent. Missing data can be ignored if it is missing at random (MAR) or missing completely at random (MCAR). In such a situation, the obvious effect lies in the reduction of number of cases for analysis. When data is missing not at random (MNAR), this may have an adverse effect on the overall result of the study and this should be of concern to a researcher.

**Reasons for Missing Data**

When any data on a variable or from a participant in a study is not present, the researcher is dealing with missing or incomplete data. Data can be missing legitimately as a result of many reasons. In a survey, one could be asked whether he/she have been engaged in teaching students with a follow-up question asking for how long. The respondent may skip the second question since he/she has never been engaged in teaching. If this is the case, it is legitimate to skip the second question. In many large datasets, it is highly possible to encounter many cases of legitimately missing data, and so a researcher needs to be thoughtful about handling this issue appropriately especially when it comes to data dressing or cleaning. In the case of legitimate missing data, "missingness" may be meaningful. The missing data in this situation informs and reinforces the status of a particular individual and can even provide an opportunity for checking the validity of an individual's responses.

Many cases of missing data are illegitimate. Calibration instruments could malfunction or fail without the notice of the researcher. Some female respondents in a survey often skip some questions especially on age by either under-stating or skipping such columns purposely because it may conflict with their official age declaration or infringe on their personal life. Other forms of illegitimate missing data may occur during data cleaning and this should be of concern to researchers because it has the potentials of introducing bias in analysis results. All methods of handling missing data go along with assumptions about the nature of the mechanism that is responsible for the missingness. Incidentally, many researchers do not appear to explicitly deal with this issue, despite knowing the obvious effect it has in substantially skewing the results(Cole, 2008).

**Types of Missing Data**

Three types of missing data have been described by little and Rubin (1989).These include: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). Identification of the types of missing data and the mechanism underlying the missingness is important for the researcher in order to properly understand how to deal with them.

**MCAR**

This is a situation where the data is missing both at random, and observed at random. MCAR situation means that the data was collected randomly, and does not depend on any other variable in the dataset. This pattern of MCAR occurs when missing values on a variable (x) are not dependent on any values or any measured variables including the variable (x) in the dataset. Simply, the observed values are a random subset of the theoretically complete dataset (Rubin, Witkiewitz, Andre & Reilly, 2007).

**MAR**

The term 'missing at random' (MAR) is a contradiction, as the missing data is anything but missing at random. The intuitive meaning of this term is better suited to the term MCAR. What MAR means is missing, but conditional on some other 'X-variable' observed in the data set, although not on the 'Y-variable' of interest(Schafer, 1997).Missing at random (MAR) implies that the data are not MCAR, and that some information as to why the data are missing is known and is examinable by other collected variables. Precisely, the pattern of MAR occurs when missing values of a variable (x) is related to other measured variables but the missing data is not a product of the variable (x) itself (Rubin, et, al, 2007). With an MAR pattern, the variables with missing data can be predicted from other acquired measures using regression equation. When data is MCAR or MAR, the reasons for missingness can be ignored because the data is missing and does not create room for much bias in our study results. However, whether the type is MAR or MCAR, it is important to note that these situations create the problem of reducing the sizes of our sample data and invariably the degree of freedom for analysis.

**NMAR**

Not Missing at Random, (or informatively missing, as it is often known) occurs when the missing ness mechanism depends on the actual value of the missing data. This is the most difficult condition to model for. If the pattern of missingness is in some way related to the outcome variables, then the data are said to be MNAR. Unlike MAR, MNAR is not predictable from other variables but rather is only explainable by the variable on which missing data exists. Obviously, a MNAR pattern results when missing values on a variable (x) is related to the values on (x).However, data missing *not* at random (MNAR) could potentially be a strong biasing influence (Rubin, 1976).

**Deletion methods**

Deletion techniques are the traditional methods of handling missing data. In case wise or list wise deletion as it is sometimes called, cases with missing values are discarded from the analysis. By this method, only cases without missing values are included in the analysis. As good as this method appears, some important data are erroneously excluded especially for datasets that include a large proportion of missing data or variables. This tends to reduce the total sample size as well as the power of significance tests (Baraldi and Enders, 2009). Similar to list wise deletion is another technique known as pair wise deletion. By this method, incomplete cases are removed on an analysis-by-analysis basis in such a way that any given case may contribute to some analysis but not the others. This method minimizes the number of cases deleted. However, it produces bias as in list wise cases since the deletion technique assumes the data to be missing completely at random (MCAR).

**Statement of the problem**

In many studies involving the use of quantitative data, researchers are often faced with challenges involving missing or incomplete data. The reasons for this are traceable to participants out rightly refusal to provide answers to questions or carelessness in handling of data during recording or data extractions. Bearing in mind that the process of developing instrument for data collection is sometimes cumbersome, a researcher should be cautious in the way and manner of handling missing or incomplete data since failure to do so can adversely affect the results of findings. Faced with this form of problem, how then can a researcher handle cases with missing data without resorting to include only cases with complete information and deleting those with missing data list wise or case wise as is most practiced especially when statistical packages are used for analysis? Conscious of the fact that methods for analyzing missing data require assumptions about the nature of the data and about the reasons for the missing observations, how can missing data be handled in line with best practice so as not to run into the risk of obtaining biased and sometimes misleading results?

**Purpose of study**

The main purpose of this study is to investigate the effect of exclusion of missing data on candidates' overall performance as well as to ascertain which of the methods of handling missing or incomplete data reduces bias in sample statistics and offers the best practice.

**Research Questions**

1. How comparable are the effect sizes produced by the various methods of handling missing data?

2. Which missing data handling technique most under estimated the effect size between deletion, mean substitution and strong imputation in the UTME Physics scores?

3. Between the traditional and modern techniques of treating missing data, which one of them provides the best practice in estimating the population parameter?

**Literature on Reporting Practices**

The purpose of this review is to explore methods which some researchers have employed in reporting missing data practices. Empirical articles examined by Peugh and Enders (2004) from two years: 1999 and 2003 reported that the year 1999, represented a demarcation line of some sort because it was in that year that APA Task Force on Statistical Inference specifically discouraged the use of list wise and pair wise deletion, asserting that the methods were among the worst methods available for practical applications (Wilkinson & Task Force on Statistical Inference, 1999).

Pigott (2001) in discussing the use of missing and incomplete data in analysis, provided an illustration of the use of modern methods in handling missing data from an intervention study which was designed to increase students' ability to control their asthma symptoms. He opined that many researchers use ad hoc methods such as complete case analysis, available case analysis (pairwise deletion), or single-value imputation. He rather suggested that model-based methods such as maximum likelihood using the EM algorithm and multiple imputation which hold more promise for dealing with difficulties caused by missing data should be used. Pigott (2001) finally advocated that a model-based method requiring specialized computer programs and assumptions about the nature of the missing data methods are appropriate for a wider range of situations than the more commonly used ad hoc methods.

Missing data plagues almost all surveys, and quite a number of designed experiments. No matter how carefully an investigator tries to have all questions fully responded to in a survey, or how well designed an experiment is; examples of how this can occur are when a question is unanswered in a survey, or a flood has removed a crop planted close to a river. The problem is how to deal with missing data, once it has been deemed impossible to recover the actual missing values. Traditional approaches include case deletion and mean imputation; (occasionally provided as an option with some software). These are the default for the major statistical packages. In the last decade, interest has been centered on Regression Imputation, and Imputation of values using the EM (Expectation - Maximization) algorithm, both of which will perform Single Imputation. More recently Multiple Imputation has become available, and is now being included as an option in the mainstream packages. In this work, the researchers intend to look at eight different methods of imputation, and compare how well these methods perform (what happens to the means and standard deviations) under different missingness mechanisms with different amounts of missing data (Sheffer, 2002).

**Method**

This study employed an ex post facto design method with a random sample of 3,000 examinee responses from the UTME Physics. From this data, twenty percent were deleted to simulate an MCAR situation as missing data.

**Data Simulation**

In order to assess the randomness of the missing data in the UTME Physics dataset used in this analysis, the approaches suggested by Little and Rubin (2002) was adopted. It has been suggested that statistical significance tests using correlations, provide a conventional estimate of the degree of randomness in missing data studies. Significant correlations in missingness between some pairs of variables suggest that the data are MAR or MNAR. Cases with missing data were compared with those without missing data for each variable. For datasets which are missing completely at random (MCAR), there was no significant difference between the two groups $F(1, 1223.42, p<.000, \eta^2 =.326)$.Also, the correlation of data missingness for each pair of variables was carried out and as expected, this was uncorrelated for data MCAR. The last randomness test which was performed used ANOVA to ascertain if those with missing data on one variable are significantly different in their aggregate score variable included in this study for purposes of comparability. From Table 1.1, it appears that there were no significant differences in the aggregate scores between those with missing data on Physics scores and those with valid scores. To generate data with mean substitution (MS), this was simply carried out by replacing missing values with the mean of the available data for that variable. In this case, the mean of Physics scores (sample 3,000). To simulate data on Multiply-imputation, SPSS version 20.0 was used which imputes missing values by using a regression technique.

**Data Simulation**

**MNAR-Low**

In simulating the MNAR-low missing data type, cases below the 20th percentile on the Physics test scores were given an 80% chance of being randomly labeled as missing, while cases between the 20th and 50th percentile on the Physics test were given a 50% chance of being randomly labeled as missing, and those above the 50th percentile had less chance of being labeled as missing on the Physics test. This situation gave high performing students better chance to respond to an item than the less performing ones. The result is that MNAR-low condition produced an increasingly biased estimate of average mean performance as well as undervaluing the SD as a result of less dispersion at the lower extreme of the distribution. Table 1.1 shows that there were substantial mean differences in aggregate achievement score between those with missing scores and those with valid Physics scores.

**MNAR-Unsafe**

Simulation of the MNAR-Unsafe situation was by giving students below the 20th percentile and above the 60th percentile on the Physics test an 80% chance of being randomly recognized as missing on the Physics test. This suggests that the missing cases were mostly at the two extremes. This should have the effect of increased nonrandom missingness without substantially skewing the population

average estimates. In doing this, the highest and lowest 30% of the students were more probable of being missed than those in the middle echelon. The resulting distribution closely matched the mean of the original population and with this reduced variance; this does not show any significant difference in the resultant dataset in terms of missingness. Table 1.1 shows that the average for MNAR-Unsafe closely estimates the population mean and under-approximates the standard deviation. This situation however, produces a non-significant correlation.

**Mean Substitution (MS)**

Many researchers are of the opinion that mean substitution is a viable, or even progressive, method of dealing with missing data. The idea some researchers used in supporting this method is that in an event where there is no data, the mean stands as the best single estimate of any score. However, this may not yield the best result since using the mean in substitution in a situation where a large number of such data is missing, inflates the effect of the missing data and thereby reducing the variance of the variable. In Table 1.1, the standard deviations were underestimated under MNAR situations. Although using mean substitution was alright in this study sample, but the effect would have been more substantial if larger percentage of the sample were missing thereby generating a more erroneous population estimates.

**Multiple-Imputation**

Multiply-imputed datasets were generated using the missing data command in SPSS version 19.0 which imputes data through a regression technique. Imputation is a method to fill in missing data with plausible values to produce a complete data set. A distinction may be made between *deterministic* and random imputation methods. While it is true that a selected sample of deterministic method always produce the same imputed value for units with the same characteristics, a random (or Stochastic) method may produce different values according to Durrant (2005). Usually, imputation makes use of a number of auxiliary variables that are statistically related to the variable in which item nonresponse occurs by means of an *imputation model* (Schafer, 1997). The main reason for carrying out imputation is to reduce nonresponse bias, which occurs because the distribution of the missing values, assuming it was known, generally differs from the distribution of the observed items. Using imputation method offers better result than deletion method in the sense that imputation allows the creation of a balanced design such that procedures used for analyzing complete data can be applied in many situations.

## Analysis and Results

### Research Question 1

1. How comparable are the effect sizes produced by the various methods of handling missing data?

When data is in MCAR situation, the estimated results are often unbiased but under the MNAR conditions the probability of misestimating is high and this can result to errors(Stuart, Azur,

Frangakis & Leaf. (2009). A look at table 1.1 shows the correlation coefficients and the associated effect sizes (Variances accounted for) computed for the original data, MCAR, MNAR-Low, MNAR-Unsafe, Mean Substitution, Strong Imputation and Multiple Imputation. From the table, the correlation coefficient for the original was p = .623 with an effect size of 38.8%. Observe that in Table1.1 the correlation coefficient is almost the same with the original data with for MCAR situation. However the result was not the same for MNAR estimate involving MNAR-Low and MNAR-Unsafe. The correlation coefficients and their effect sizes were .571, 32.604% and .378, 14.288 respectively. In these MNAR situations, the relationship between the UTME Physics scores and the aggregate scores represented by the correlation coefficient begins to exhibit diminishing returns when compared with the original data. The idea behind the Mean Substitution is that instead of leaving the variable empty without value, the mean of the UTME Physics is the best estimate of the candidate score. From Table 1.2, the correlation coefficient of MCAR is .590 with the effect size of 34.81% while the correlation coefficient and the effect sizes of the MNAR-Low and MNAR-Unsafe is .491, 24.108% and .478, 22.848% respectively. Observe that the variance accounted for was much lower in MCAR and MNAR-Low for Mean Substitution when compared with same situation in the original data set. This means that using mean substitution in handling missing data can arbitrarily create inaccurate estimate.

The effect of this Mean Substitution appears more vivid under MNAR-Low when we look at the Mean and Standard Deviation of the Physics scores. This is because the missing data are likely to be from the low performing students and in the case the mean is a poor estimate of their performance. Hence the mean score was over estimated with the mean 52.02 which also under estimated the standard deviation by 6.958. The Mean Substitution therefore, under estimated the standard deviation approximately by 30.3%.

In the case of strong imputation method of handling missing data, the population parameters are closely replicated for the MCAR, MNAR-Low and MNAR-Unsafe. This means that the issue of under estimation or over estimation is significantly reduced. This result is better when compared with the results Mean Substitution. However, the standard deviation was more under estimated in MNAR-Unsafe. In the case of Simple Imputation, the effect size of MCAR is almost the same with that of strong imputation. However the population means of MNAR-Low and MNAR-Unsafe were overestimated while the standard deviation was underestimated.

**Table 1.1: Summary of Effects of Missing Data Corrections on UTME Physics Scores**

| | N | Mean Physics Score | SD Physics Score | Kurtosis, Skew of Physics Score | Mean UTME Aggre-gate Score (Not Missing) | Mean UTME Aggre-gate Score (Missing) | F-ratio | Average Error of Estimate (Mean) | Corre-lation with Aggre-gate Score (r.) | Effect Size ($r^2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| Original Data | 2,940 | 50.02 | 10.096 | 2.476, -1.569 | | | | | 0.623* | 38.8 |
| Missing Completely at Random (MCAR) | 2,644 | 50.04 | 10.042 | 2.489, -1.570 | | | < 1,ns | 0.195 | .623* | 38.80 |
| Missing Not at Random (MNAR-Low) | 2,538 | 51.41 | 8.319 | 5.881, -1.859 | 178.31 | 180.12 | 1223.42, p<.000, $\eta_p^2$ =.326 | 0.165 | .571* | 32.60 |
| Missing Not at Random (MNAR-Unsafe) | 2,157 | 52.52 | 5 | 8.441, -.501 | | 181.08 | 358.012, p< .000, $\eta_p^2$=.143 | 0.108 | .378* | 14.29 |

| | N | Mean Physics Score | SD Physics Score | Kurtosis, Skew of Physics Score | Mean UTME Aggre-gate Score (Not Missing) | Mean UTME Aggre-gate Score (Missing) | F-ratio | Average Error of Estimate (Mean) | Corre-lation with Aggre-gate Score (r.) | Effect Size ($r^2$) |
|---|---|---|---|---|---|---|---|---|---|---|
| **Mean Substitution** | | | | | | | | | | |
| | | | | | | | | | | |
| | | | | | | | | | | |
| MCAR | 2,996 | 50.07 | 9.982 | 2.519, -1.571 | | | | 0.182 | .590** | 34.81 |
| MNAR - Low | 2,996 | 52.02 | 6.958 | 7.356, -1.807 | | | | 0.127 | .491** | 24.11 |
| MNAR - Risky | 2,996 | 50.91 | 6.294 | 10.420, -2.162 | | | | 0.115 | .478** | 22.85 |
| **Strong Imputation** | | | | | | | | | | |
| MCAR | 2996 | 50.02 | 10.001 | 2.580, -1.583 | | | | 0.183 | .622** | 38.69 |
| MNAR - Low | 2,432 | 49.85 | 8.754 | 1.332, -1.298 | | | | 0.178 | .563** | 31.70 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| MNAR-Unsafe | 2,389 | 51.74 | 5.646 | (-.298,-0.674) | | | | 0.116 | .426** | 18.15 |
| **Weak Imputation** | | | | | | | | | | |
| MCAR | 2,202 | 50.12 | 9.865 | 2.656, -1.588 | | | | 0.21 | .613** | 37.58 |
| MNAR-Low | 1,815 | 55.16 | 6.759 | 17.226, -3.860 | | | | 0.135 | .497** | 24.70 |
| MNAR - Risky | 1,713 | 55.82 | 5.608 | 27.008, -4.560 | | | | 0.158 | .428** | 18.15 |

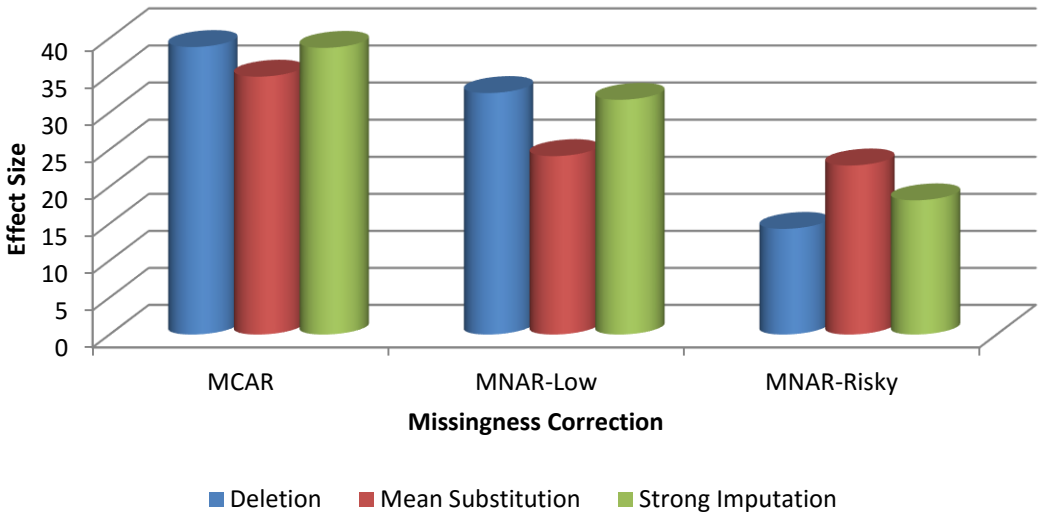
**Research question 2:** Which missing data handling technique most under estimated the effect size between deletion, mean substitution and strong imputation in the UTME Physics scores?

Figure 1 depicts the relative under estimation of effect sizes when data are treated using deletion, mean substitution and strong imputation techniques. Under MCAR condition, the effect sizes when missing data are treated with deletion and strong imputation methods appear similar, but mean substitution method created a more inaccurate population estimate. Mean substitution which allows substituting identical scores especially for large portions of the sample falsely reduced the variance of the UTME Physics variable and so with more data missing, the effect size reduced further.

Under MNAR-low situation, the mean substitution method drastically reduced the variance thereby creating more inaccurate population estimates than the deletion and strong imputation methods when data are treated under the MNAR situation. This results in the making the effect size smaller because substantial proportions of the sample were missing but were simply replaced substituting the variable with identical scores which may be likely that of low performing students and which may not be a true estimate of their performance.

With data treated under the MNAR-Unsafe, the whole picture appears to be more affected. Strong imputation method which is regression-based and has strong predictive value with aggregate UTME scores uses information available in the existing data to estimate a better value. The mean and standard deviation appears to be closely replicated under the MCAR, MNAR-low and MNAR-Unsafe situation. Over estimation or under estimation appears to be significantly reduced and with population estimates more closely approximated than the mean substitution and deletion methods. The mean substitution method under the MNAR-Unsafe showed a better estimation than the deletion and strong imputation methods. This result can be misleading because the scores used in substituting the missing values may have been artificially increased.

**Figure 1: Under Estimation of Effect Size When Missing Data are Treated Through Deletion, Mean Substitution and Strong Imputation**




**Research question 3:** Between the traditional and modern techniques of treating missing data, which one of them provides the best practice in estimating the population parameter?

Table 1.3 provides a summary of traditional and modern methods of handling missing data under the MNAR-Unsafe situation. From the table, it can be seen that the apart from the original dataset under the MNAR situation, the mean estimates were overestimated while the standard deviations were under estimated. The correlation coefficients obtained by relating the Physics scores with the aggregate UTME scores provided much less effect sizes than the multiple imputation method which represents one of the modern methods of handling missing data. For instance, under the traditional method, strong imputation technique produced population parameter estimates of 51.74 mean values and a standard deviation of 5.646.

The multiple imputations (MI) method which represents the modern method of treating missing values uses a more advanced technique such as the Expected Maximization (EM) algorithm, Maximum Likelihood Estimation (MLE) or the Markov Chain Monte Carlo (MCMC) simulation method to estimate missing values. Procedure for correcting missing data in MI involves creating multiple versions of the same dataset. Results produced using MI method with EM algorithm appears to be more consistent than any of the traditional methods of correcting missing values. However, when the proportion of missing data is reasonably small, missing values estimated using strong imputation appears to be plausible and comparable with MI, despite the fact that it sometimes over-fits the data (Osborne, 2008). For this reasons mentioned, MI has the potential of more accurately estimating the population parameter than any of the traditional methods and therefore offers the best practice in handling of missing data.

**Table 1.3: Summary of Traditional and Modern Methods of Handling Missing Data with Faculty of Study as Predictor in MNAR -Extreme Situation**

| | N | Mean Physics Score | SD Physics Score | Kurtosis, Skew of Physics Score | Correlation with Aggregate Score (r.) | Effect Size ($r^2$) |
|---|---|---|---|---|---|---|
| **Traditional Methods** | | | | | | |
| Original Data | 2,940 | 50.02 | 10.096 | -1.569, 2.476 | .623** | 38.810 |
| Complete Case Analysis | 2,159 | 52.52 | 5.000 | -.501, 8.441 | .378** | 14.288 |
| Mean Substitution | 2,996 | 50.91 | 6.294 | -2.162, 10.420 | .487** | 22.484 |
| Strong Imputation | 2,389 | 51.74 | 5.646 | -.298, 0.674 | .426** | 18.147 |
| Weak Imputation | 1,713 | 55.82 | 5.608 | -4.560, 21.008 | .428** | 18.318 |
| **Multiple Imputation Using EM Algorithm of SPSS (Modern method)** | | | | | | |
| Imputation_1 | 2996 | 49.859 | 10.315 | -1.634, 2.780 | .532** | 28.302 |
| Imputation_2 | 2996 | 49.869 | 10.316 | -1.643, 2.851 | .518** | 26.832 |
| Imputation_3 | 2996 | 49.858 | 10.316 | -1.635, 2.785 | .522** | 27.248 |
| Imputation_4 | 2996 | 49.854 | 10.367 | -1.664, 2.961 | .518** | 26.832 |
| Imputation_5 | 2996 | 49.863 | 10.303 | -1.633, 2.789 | .526** | 27.667 |

## Conclusions

When exploring missing data, it is important to understand the reasons why data is missing or the mechanism of missingness. When data is MAR or MCAR, non-response by participants can be ignored. However, whether MAR or MCAR, missing value conditions in variables is of concern to the researcher because data is randomly missing and this condition often affects the sample size or the degree of freedom for an analysis. The non-ignorable missing value conditions in variables which is characterized by MNAR situation potentially introduces a strong biasing influence in an analysis (Rubin, 1976). In the same way, the traditional way of missing value treatment involving deletion pairwise or list wise as well as mean substitution changes the sample being analyzed. This change however, depends on the variables involved in the analysis. This can be a problem when it comes to replicating the data which invariably increases the odds of error of inference (Schafer & Graham, 2002). Case deletion is therefore not encouraged because it destroys the variance structure within the data.

Strong and weak imputation methods can produce encouraging results if the amount of missingness is below 10% and under the MAR or MCAR situations when the focus is on the mean, but if the interest is on the variance structure, the two methods should not be used. Results of MI method can

be promising even with up to 25% missing values because even though no method can reproduce the population parameters, estimation by MI has been shown to produce a more reliable estimate of the population parameters as well as a more steady skewness and kurtosis.

**Recommendations**

Resolute effort should be made when cleaning data in order to understand the nature of the data and the mechanism of missingness to enable the researcher make appropriate decision as to which form of data correction method to adopt.

Avoid the deletion method unless the data is one with legitimate missing values which makes it MCAR. Do not attempt to use weak imputation method even when the data is MCAR.

In collecting data for analysis, the research should try as much as possible to avoid situations where participants would arbitrarily ignore response to particular items. In the event of this becoming unavoidable and the percentage of missing data is reasonably not high, multiple imputation should be used in correcting the missing values.

**Reference**


Baraldi, A. N. & Enders, C. K. (2009). An introduction to modern missing data analyses. Arizona State University, United States. *Journal of School Psychology 48 (2010) 5–37*

Durrant, G. B. (2005). Imputation Methods for Handling Item-Nonresponse in the Social Sciences: ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute (S3RI), University of Southampton. NCRM Methods Review Papers NCRM/002

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data.* New York: Wiley.

Little, R.J.A., (1988). A Test of Missing Completely at Random for Multivariate Data with Missing Values. *Journal of the American Statistical Association* **83** (404) 1198-1202

Meng, X-L, (1994) Multiple Imputation with Uncongenial Sources of Input. *Statistical Science* **9** 538-573

Osborne, J. W. (2008). Creating valid prediction equations in multiple regression: Shrinkage, double cross-validation, and confidence intervals around prediction. In J. W. Osborne (Ed.), *Best practices in quantitative methods.* (pp. 299–305). Thousand Oaks, CA: Sage.

Peugh, J. L. & Enders, C. K. (2004).Missing data in educational research: a review of reporting practices and suggestions for improvement. *Review of educational research, 74, 525-556*

Pigott, T. D. (2001). A Review of Methods for Missing Data. Educational Research and Evaluation 1380-3611/01/0704-353$16.00 2001, Vol. 7, No. 4, pp. 353±383 # Swets & Zeitlinger,.Loyola University Chicago, Wilmette, IL, USA

Rubin, D. (1976). Inference and missing data. Biometrika, 63(3), 581-592.

Rubin, L. H., Witkiewitz, K., St. Andre, J. & Reilly, S. (2007)Methods for Handling Missing Data in the Behavioral Neurosciences: Don't Throw the Baby Rat out with the Bath Water .*The Journal of Undergraduate Neuroscience Education (JUNE), Spring 2007, 5(2):A71-A7.*

Schafer, J. (1997). *Analysis of incomplete multivariate data*. London: Chapman & Hall/CRC.

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

Scheffer, J. (2002). Dealing with missing data. *Res. Lett. Inf. Maths. Sci. (2002) 3, 153-160. I.I.M.S Quad A, Massey University Auckland, 1310.*

Stuart, E. A., Azur, M., Frangakis, C., & Leaf, P. (2009). Multiple imputation with large data sets: A case study of the children's mental health initiative. *American Journal of Epidemiology, 169*(9), 1133–1.

Wilkinson & Task Force on Statistical Inference (1999, abbreviated as the 1999 APA Task Force Report).