Kwantlen Psychology Student Journal
Issue 6, August 2024
Kwantlen Polytechnic University

# Multiple Imputation:
# An Investigation of the Missing Data Techniques Effectiveness

Melissa E. M. Campion[1]

## Abstract

Multiple Imputation (MI) is one of the most reliable techniques in addressing missing data due to partial or incomplete responses from a portion of the sample. MI has been particularly useful when handling missing data patterns such as Missing Completely at Random (MCAR) and Missing at Random (MAR). However, there have been some debates on its use when it comes to the Missing Not at Random (MNAR) pattern due to the bias it creates. This paper further examines the complexities of using MI to accurately complete missing data sets, exploring both its effectiveness and limitations.

*Keywords:* multiple imputation, missing data, data analysis

[1] melissa.campion@student.kpu.ca; Written for Psychology Quantitive Data Analysis using SPSS (PSYC 4900). Thank you, Dr. Minosky, for the recommendation and assistance in publishing this paper.

# Multiple Imputation:
# An Investigation of the Missing Data Techniques Effectiveness

Multiple Imputation (MI) is a powerful method for handling missing data by aggregating a series of single imputations. Methods that handle missingness, such as MI, are important to consider, as missing data is a common problem frequently occurring in scientific research (van Ginkel et al., 2020). These methods, or missing data techniques, are employed by analysts when dealing with item non-response—a type of missingness that happens when a portion of the sample provides incomplete, partial responses (Buhi et al., 2008; Garson, 2019). These unreported values represent data that would have been meaningful for the analysis if included. Therefore, the missingness can negatively influence parameter estimates, internal validity, sample size, statistical power, and the probability of biased results (Blankers et al., 2010; Little & Rubin, 2020; Pepinsky, 2018). Missing data and missing data techniques are a massive topic with various components, options, strategies, and theories. In addition to this, there is considerable variation within the MI technique itself. This paper critically analyzes the MI regression approach by exploring the factors involved in the MI process and the various arguments proposed in the literature that support and criticize the use of MI.

## Missingness Patterns

Since missing data is a form of measurement error that can distort and bias analyses, understanding the underlying mechanisms (or patterns) behind the missingness is essential (Garson, 2019; van Ginkel et al., 2020). These mechanisms are the causes in which the participant information becomes “lost” or missing (Buhi et al., 2008). The awareness and identification of these mechanisms are vital in deciding the best procedure to handle them (Buhi et al., 2008; Little & Rubin, 2020; Pepinsky, 2018). There are three patterns of missingness to be aware of: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

The first pattern, MCAR, happens when the probability of the missingness is unrelated to any of the reported or missing data (Buhi et al., 2008). Essentially, this missingness occurs entirely by chance and has no relation to the study’s topic, design, procedures, or measures (Buhi et al., 2008). These values are a completely random subset of the sample; when omitted, the remaining participants reflect an equally representative sample as the initial (Garson, 2019; Pepinsky, 2018; van Ginkel et al., 2020). MCAR happens, for example, when a question on a survey is skipped by accident (van Ginkel et al., 2020). Although considered the ideal because of the ease required to

handle MCAR, the chance that missing data is genuinely random is rare (Blankers et al., 2010; Garson, 2019).

The name of the second pattern, MAR, is misleading because the probability of this missingness is conditional and can be predicted by the reported data and, thus, not exactly random (Blankers et al., 2010; Pepinsky, 2018). In other words, the probability of missing data is a function of the reported data and is traceable (Buhi et al., 2008; van Ginkel et al., 2020). For example, women may be less likely than men to report their weight (Buhi et al., 2008). If this is the case, then the probability of the weight value missing depends on the reported gender value, and thus, the missingness is explained by whether the participant is a man or woman and not by how much they weigh (Buhi et al., 2008). The predictive capability of MAR allows for the completion of missing datasets based on participants' completed data (Jakobsen et al., 2017). Both MAR and MCAR are considered ignorable missingness, which means the reason(s) the data is missing can be ignored, and a missing data technique can be used to manage the problem (Buhi et al., 2008). MAR is the most common type of missingness and is manageable through various missing data techniques, including MI (Garson, 2019).

The last pattern, MNAR, accounts for missingness caused by bias or systematic influences (Buhi et al., 2008; van Ginkel et al., 2020). In other words, the probability of missingness depends on unreported data and cannot be predicted by the reported data as it depends on the missing data itself (Pepinsky, 2018). An example would be overweight participant's decreased likelihood to report their weight (Buhi et al., 2008). In this situation, the probability of missing weight values relies solely on the participants' weight, which is not reported (Buhi et al., 2008). This pattern is considered a non-ignorable missingness as the analysis cannot account for MNAR data (Garson, 2019). MNAR is the most problematic pattern to detect and address because no standard missing data techniques can remove the bias MNAR creates (Blankers et al., 2010; Buhi et al., 2008; Cook, 2021; Pepinsky, 2018).

**Missing Data Techniques**

Statisticians have developed numerous missing data techniques to handle missingness (Buhi et al., 2008). The primary goal of missing data techniques is to obtain the most unbiased estimates (Blankers et al., 2010). However, researchers must choose a missing data technique that best fits their data to get the most accurate estimated parameter while avoiding error inflation (Blankers et al., 2010). Standard error, the estimate of how well the data represents the population,

requires special attention because increases in this value reduces statistical power, increasing the chance of Type II error (Garson, 2019). Two categories of missing data techniques are well known: deletion and imputation.

Deletion Techniques, such as listwise and pairwise deletion, are common, user-friendly, and create complete datasets (Buhi et al., 2008). However, these techniques are not often recommended as they have many disadvantages. Deletion techniques are wasteful because they remove valuable information from the data, reducing the sample size, statistical power, and generalizability (Buhi et al., 2008; van Ginkel et al., 2020). Listwise deletion (complete case analysis) excludes all cases with missing values (Blankers et al., 2010; Garson, 2019; van Ginkel et al., 2020). Pairwise deletion is similar but removes variables with missing values (not entire cases) so that all the available data and cases are used (Buhi et al., 2008; Cook, 2021). Pairwise is less wasteful, but removing variable-by-variable data makes the final sample size unclear, thus complicating the standard error calculation (van Ginkel et al., 2020).

Imputation techniques replace each missing value with a reasonable guess or estimate and run the analysis as if the dataset were complete from the beginning (Buhi et al., 2008; Little & Rubin, 2020; van Ginkel et al., 2020). Ultimately, the missing data techniques overcome item-nonresponse by estimating a subject's response to create a best-guess replacement value (Garson, 2019). Single imputation resolves the wastefulness problem but introduces additional bias in statistical analysis (van Ginkel et al., 2020). These missing data techniques are general and flexible, but imputing a single value creates biased, unreliable results (Little & Rubin, 2020; McKnight et al., 2007; van Ginkel et al., 2020).

In mean imputation, each missing value is replaced with the variable mean (van Ginkel et al., 2020). There is an overwhelming consensus in the literature to avoid this missing data technique due to its negative consequences on variability (Buhi et al., 2008). This missing data technique artificially reduces variance (especially standard error) and confidence intervals (Blankers et al., 2010; Buhi et al., 2008; van Ginkel et al., 2020). Regression imputation is another single imputation missing data technique that replaces the missing data with values predicted from a linear regression model. This process creates bias by artificially decreasing variance and increasing covariance (van Ginkel et al., 2020). An extension of this method is stochastic regression imputation. Like regression imputation, this missing data technique uses a regression model to predict missing values but improves the estimates by adding a randomly generated error

term to reflect the uncertainty of the predicted value (Little & Rubin, 2020; Murray, 2018; van Ginkel et al., 2020). While including the error term reduces the biases the other two imputation techniques create, the analyses following stochastic regression imputation treat the imputed values as real data (van Ginkel et al., 2020). More simply, the assumed certainty of the data biases results such as *p* values and confidence intervals (van Ginkel et al., 2020). Ultimately, unless missingness is MCAR, all these missing data techniques introduce bias into the analysis and artificially suppress variance, subsequently reducing statistical power and increasing the chances of Type II error (Buhi et al., 2008; van Ginkel et al., 2020).

**Multiple Imputation Theory**

As previously mentioned, MI is a sophisticated and effective missing data technique that combines numerous single imputations to deal with missing data. MI uses reported data in an iterative process to create a series of simulated datasets with different values in each (Buhi et al., 2008; McKnight et al., 2007; Pepinsky, 2018). These values are created from linear regression models when the variables are continuous and logistic regressions when categorical (van Ginkel et al., 2020). MI is a three-step process: imputation, analysis, and pooling.

The first step is imputation. Just as a single imputation generates one set of plausible values for a data set, MI generates multiple sets (Jakobsen et al., 2017). This process is accomplished through numerous regression imputations that create new, complete datasets by taking a random subset of the data and using it to conduct a linear regression model (Cook, 2021; van Ginkel et al., 2020). The regression model produces a line of best fit and prediction equation ($Y = a + bX$). Since MI typically assumes the data is MAR, the missing values are predicted by reported data values, expressed through a unique regression equation (Missing Value = a + bReported Value). However, like stochastic regression, MI improves the estimated values by adding a randomly generated error to each predicted value (Missing Value = a + bReported Value + e). This equation is then used to predict an estimated value for each missing value. These estimated values are combined with the original data to create a new complete dataset. This step results in a series of new datasets with all the same reported values, each containing its unique set of estimated values that replace what was once missing (Buhi et al., 2008).

In the analysis step, the new imputed datasets are analyzed through a standard statistical procedure (i.e., t-test, ANOVA, regression) to produce slightly different parameter estimates and

standard errors for each (Buhi et al., 2008; van Ginkel et al., 2020). This step mirrors a regular analysis typically conducted, except the analysis is performed separately for each new dataset.

Finally, the pooling step aggregates the estimated parameters and standard errors from each imputed dataset to create a single, best estimate (Buhi et al., 2008; Gorard, 2020; McKnight et al., 2007). These aggregated results account for the uncertainty created by the imputation process and subsequently incorporate it into the final pooled results (Cook, 2021; Murray, 2018; van Ginkel et al., 2020). Accounting for the uncertainty in the estimated variance is calculated by adding the within-imputation variance—an estimate of the variance if the data had been complete—and the between-imputation variance—an estimate of the excess variance due to the missing values (Murray, 2018). Accounting for this variability (or uncertainty) sets this missing data technique above the others.

**Support for MI**

Many researchers who support using MI argue that the missing data technique provides the most accurate, unbiased, and valid results (Blankers et al., 2010; Buhi et al., 2008; Cook, 2021; van Ginkel et al., 2020). MI has consistently proven effective in providing generalizable estimates and recovering the population variance critical in statistical inferences (McKnight et al., 2007; Pepinsky, 2018). This missing data technique has been tested through simulation studies that show MIs superiority in restoring parameter estimates, confidence intervals, and statistical significance close to the original values (Blankers et al., 2010; Buhi et al., 2008; van Ginkel et al., 2020).

Another common argument favouring the missing data technique is MIs ability to account for random error (e.g., Blankers et al., 2010; Buhi et al., 2008; Cook, 2021; McKnight et al., 2007). MIs unbiased accuracy stems from the random error added to account for the uncertainty of the replacement values. This adjustment allows MI to estimate how much the missing and imputed values influence the parameter estimates and the resulting statistical conclusions (McKnight et al., 2007). This process reduces the likelihood of a Type I error by increasing the variability used to adjust the standard error (McKnight et al., 2007).

In addition to these two major advantages, minor arguments supporting MI use are plentiful. First, MI solves the wastefulness problem in deletion techniques by avoiding removing valuable information (van Ginkel et al., 2020). Next is the flexibility of MI. Once the imputed datasets are created, nearly all analyses can be conducted on the data (Buhi et al., 2008). MI's ability to handle missingness in categorical variables is also an advantage (Buhi et al., 2008).

Finally, the superiority of MI has been attributed to the benefits of using the missing data technique with longitudinal studies, a design notorious for missingness (Buhi et al., 2008).

### Critiques of MI

Although theoretically considered optimal, many applied researchers are reluctant to use MI (van Ginkel et al., 2020). The biggest argument against using MI is the lack of clear guidelines (Cook, 2021). This lack of consensus is apparent in the recommended number of imputations and missingness parameters. The suggested imputations in the literature range from 3–5 (Garson, 2019), 3–10 (Blankers et al., 2010; Cook, 2021; McKnight et al., 2007), 5–20 (Buhi et al., 2008; Garson, 2019), 5–50 (Jakobsen et al., 2017), and 2–100 (Garson, 2019). The recommendations for missingness parameters are mixed. Researchers tend to agree that MI and most other missing data techniques are irrelevant when the missingness is under 5% (Buhi et al., 2008). However, the maximum amount of missingness varies between 20% (Garson, 2019), 30% (Jakobsen et al., 2017), 40% (Cook, 2021; Jakobsen et al., 2017), and 50% (Blankers et al., 2010; Buhi et al., 2008; Garson, 2019). The lack of firm MI guidelines may cause extreme limitations as incorrectly applying the technique can create untrustworthy results and more bias than initially started (van Ginkel et al., 2020).

Another MI critique is the use of pooled data when analyzing results. Researchers argue against creating and analyzing what they deem fake data instead of authentic reported data (Cook, 2021). This view causes some to question how valid MI results are due to the numerous statistical adjustments that transform the dataset differently than the original reported value (Buhi et al., 2008).

The final argument against using MI surrounds the appropriateness when missingness is MNAR. Since MI assumes that missingness is MAR, the technique uses the reported values to predict the missing values. If missingness is MNAR, the unbiased results are no longer guaranteed (Cook, 2021; van Ginkel et al., 2020). However, as previously stated, most missing data techniques end up with biased results when missingness is MNAR.

## Conclusion

While no missing data technique is perfect, many researchers agree that MI is the best method for handling missing data. After analyzing the support and critiques of MI, I found that the advantages of using MI outweigh the disadvantages. While MI use is rising, the method is still far from being the standard in statistical analyses (McKnight et al., 2007; van Ginkel et al., 2020).

Misunderstandings and misconceptions of procedures cause researchers to instill misguided distrust for the missing data technique (van Ginkel et al., 2020). van Ginkel et al.'s (2020) article addresses many misconceptions regarding MI through theoretical and practical argumentation. Simulation studies have also confirmed MI's superiority over deletion and single imputation (Blankers et al., 2010; Buhi et al., 2008; van Ginkel et al., 2020). Regardless of the missingness pattern, MI is always preferred over deletion: under MCAR, it produces more statistical power, MAR produces more power and unbiased results, and MNAR produces less biased results than deletion (van Ginkel et al., 2020). MI is more complex than basic missing data techniques but produces the most accurate results (Blankers et al., 2010).

## References

Blankers, M., Koeter, M. W. J., & Schippers, G. M. (2010). Missing data approaches in eHealth research: A simulation study and a tutorial for nonmathematically inclined researchers. *Journal of Medical Internet Research, 12*(5), 54. https://doi.org/10.2196/jmir.1448

Buhi, E. R., Goodson, P., & Neilands, T. B. (2008). Out of sight, not out of mind: Strategies for handling missing data. *American Journal of Health Behavior, 32*(1), 83–92.

Cook, R. M. (2021). Addressing missing data in quantitative counseling research. *Counselling Outcome Research and Evaluation, 12*(1), 43–53. https://doi.org/10.1080/21501378.2019.1711037

Garson, G. D. (2019). *Missing values analysis and data imputation*. Statistical Associates Blue Book Series.

Gorard, S. (2020). Handling missing data in numeric analyses. *International Journal of Social Research Methodology, 23*(6), 651–660. https://doi.org/10.1080/13645579.2020.1729974

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials – a practical guide with flowcharts. *BMC Medical Research Methodology, 17*(162), 1–10. https://doi.org/10.1186/s12874-017-0442-1

Little, R. J. A., & Rubin, D. B. (2020). *Statistical analysis with missing data* (3rd ed.). John Wiley & Sons. https://doi.org/10.1002/9781119482260

McKnight, P. E., McKnight, K. M., Sidani, S., & Figueredo, A. J. (2007). *Missing data: A gentle introduction*. Guilford Press.

Murray, J. S. (2018). Multiple imputation: A review of practical and theoretical findings. *Statistical Science, 33*(2), 142–159. https://doi.org/10.1214/18-STS644

Pepinsky, T. B. (2018). A note on listwise deletion versus multiple imputation. *Political Analysis, 26*(1), 480–488. https://doi.org/10.1017/pan.2018.18

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2020). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment, 102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680