EVIDENCE FOR A MULTIPLE IMPUTATION APPROACH TO MNAR MECHANISMS

A THESIS

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE

MASTER OF SCIENCE

BY

BRENDAN JAMES SHANAHAN

THESIS COMMITTEE:

DR. JOCELYN BOLIN, CHAIR

DR. HOLMES FINCH

DR. MICHAEL MACKAY

BALL STATE UNIVERSITY

MUNCIE, INDIANA

JULY 2021

Evidence for a multiple imputation approach to MNAR mechanisms

The problem of missing data is by no means a new problem. In fact, it's a problem that has really not changed since the beginning of behavioral research. What has changed, however, is the number of different ways that researchers can handle their missing data. Accounting for missing data can be tricky; there is no one-size-fits-all solution. In fact, the methodology that researchers use to handle their missing data is often specific to the conditions of the study, the reason for the missing data and the patterns of missingness that exist in the data. The conditions of the study include the methods used in collecting data and the purpose for which the data is being collected. Many researchers have examined the existing methods for handling missing data. There are many suggestions on the methods that should be used in almost any case, but there is still debate surrounding this problem. With advances in statistical research and analysis, more sophisticated methods for handling missing data have been developed.

The most popular methods being utilized are procedures based on completely recorded units and imputation methods (Little & Rubin, 2019). Unfortunately, these popular methods require certain assumptions about the nature of the missing data (i.e., that the missingness mechanism is ignorable). There exist methods which are more robust and can address some of the issues associated with the aforementioned methods (van Buuren, 2018; Galimard et al., 2016). These robust methods have been slow to migrate to the social sciences. In this study, the problem of missing data is examined and methods for handling missing data are defined. Furthermore, this study aims to examine the effectiveness of a multiple imputation by chained equations procedure using James J. Heckman's 1976 method, also called the sample selection method. This particular method has been shown to be useful in the event that some of the missing data does not meet the assumption of an ignorable missingness mechanism, meaning the

data has been determined to be missing not at random (MNAR). The results of this study will compare complete-case analysis and traditional MICE, both of which rely on the assumption of an ignorable missingness mechanism, to the method proposed by Galimard et al. (2016). Galimard et al., in their 2016 study, proposed using Heckman's model as an imputation model with a chained equations, two-step estimation process when MNAR data appears in the outcome and MAR data is present in predictors. The aim is to provide information regarding best practices for researchers in handling their missing data.

Literature Review

Missingness mechanisms

Before one can determine the method to be used in handling missing data, one must first examine two aspects of the missing data. The first being: why is the data missing, referred to as missingness mechanisms. "Missingness mechanisms are crucial because the properties of missing data methods depend very strongly on the nature of the dependencies in these mechanisms" (Little & Rubin, 2019). There are three missingness mechanisms: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). For data that are MCAR, the missingness does not depend on the values of the missing or observed data. In other words, the cause for the missing data is not related to the data itself and the probability of being missing is the same for all cases, that is, the missing data are simply a random subset of the data (van Buuren, 2018). An example of MCAR is when a participant simply skips a question on a survey by accident. When the missingness depends on some grouping defined by the observed data, the missing data is considered MAR. That is, the probability of an observation being missing is related to one or more of the other observed variables and is not related to the value of the missing variable. An example from Cheema

(2014) clarifies this definition: “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” (p. 491). Ideally then, with MAR data, one can obtain a nonbiased set of data if the method for handling the missing data incorporates, or controls for, the related variables. Both Little and Rubin (2019) and van Buuren (2018) note that this mechanism is a much less restrictive assumption and MAR is more general and more realistic than MCAR. Modern missing data methods generally start from the MAR assumption. If neither MCAR nor MAR apply, then the data is MNAR meaning the data are missing for reasons unrelated to the observed variables in the data set and perhaps unknown to us. Furthermore, the value of the missing observation has a direct relationship with its likelihood for being missing. For example, those with higher salaries are less likely to report it when asked to do so; in this case, the probability of being missing increases as the value of the outcome variable (salary) increases.

Literature from Pigott (2001) suggests that when data are MCAR or MAR, the mechanism is considered to be ignorable. MNAR data, then, are considered nonignorable. Pigott states that, with nonignorable missing data the reasons for the missing observations depend on the values of those variables. Her example is when collecting data from students with asthma about the severity of their symptoms; individuals may be more likely to miss the data collection instance because the severity of their symptoms is so great. In this case, the data is MNAR and this time of data collection is the optimal time to investigate the possibility of nonignorable missingness mechanisms. This is important because when they are ignorable, researchers can ignore the reasons for missing data in the analysis of the data. This allows the researcher to simplify the analysis considerably. Schafer and Graham (2002) use this same *ignorable* and *nonignorable* terminology when referring to missingness mechanisms. Schafer and Graham

(2002) go on to state that there is no way to test whether MAR holds in a data set, except by obtaining follow-up data from nonrespondents. Peugh and Enders (2004) concur that "MAR is an untestable assumption and could only be verified if we had knowledge of the missing [outcome] scores" (p. 527).

This reiterates Pigott's (2001) idea that the data collection phase is the optimal time to investigate possible MNAR confounds. Because there is no way to test for MAR data, Pigott (2001) suggests recording reasons for missing data as this will allow the researcher to present justification for the method selected. Pigott (2001) also states that "one strategy for increasing the probability of an ignorable response mechanism is to use more than one method for collecting important information" (p. 357). The literature on these mechanisms is very similar as these mechanisms have been widely accepted and used in the literature for many years now.

Missing data, particularly when the missingness mechanism is nonignorable, can have a great impact on subsequent analyses. Improper handling of missing data can lead to the degradation of both internal and external validity (Schafer, 1997; Finch, 2019). MCAR data can always be considered ignorable, as the missing values will not produce bias in results obtained from subsequent analyses (Pigott, 2001; Little and Rubin, 2019; Schafer & Graham, 2002). MAR data can be ignorable so long as the related, explanatory variables are controlled for or the observed data can account for or correct for the effects of the missing data (van Buuren, 2018; Anderson et al., 1983; Heitjan & Basu, 1996). As stated previously, most modern missing data methods use the MAR assumption. At the very least, maximum likelihood and multiple imputation methods require an assumption of ignorable missingness. The methods of complete-case analysis, methods that impute and model-based methods will be discussed further in this review.

Patterns of Missingness and Popular Methods

A second aspect of missing data which must be examined before selecting appropriate methods is: what are the patterns of missingness? The literature from Little and Rubin (2019) and Schafer and Graham (2002) distinguish missingness patterns; these describe which values are missing and observed in the data. This is useful as some methods of analysis are only meant for particular patterns of missing data. Missing data methods can be classified into four non-mutually exclusive categories: procedures based on completely recorded units; weighting procedures; single imputation; and model-based methods. Here, differing patterns of missingness will be discussed and appropriate, popular methods will be defined and discussed with advantages and limitations put forth in the literature. Supplemental literature will then be offered in support or in opposition.

Unit Nonresponse and Deletion Methods

Let us first define unit nonresponse in survey methods. Unit nonresponse, where "a questionnaire is administered, and a subset of sampled individuals does not complete the questionnaire because of noncontact, refusal, or some other reason" is typically addressed using complete-case or available-case analysis (Little & Rubin, 2019, p. 10). These methods are considered procedures based on completely recorded units.

Complete-Case Analysis

Procedures based on completely recorded units implement statistical analyses on only units that have been observed, dropping the units that are missing. These procedures include complete-case analysis, also called listwise deletion, and available-case analysis, also called pairwise deletion. Complete-case analysis involves simply dropping the cases with missing data and analyzing only units with completely recorded data. Advantages of this procedure are that it

is a simple, quick procedure. Although Little and Rubin admit that it may be satisfactory for small amounts of missing data, it is generally not very effective for drawing inferences in populations and can lead to serious biases.

Essentially, analysis loses precision when cases are dropped and bias occurs when data is not MCAR (Ayilara et al., 2019). The degree of these disadvantages depends on the fraction of complete units, pattern of missingness, the extent to which complete and incomplete units differ and the estimands of interest (Little & Rubin, 2019). Little and Rubin (2019) suggest that to test whether the complete units are possibly a random subsample of the original sample (meaning the MCAR assumption is reasonable) one can compare the distribution of a particular variable based on complete units with the distribution of the same variable based on incomplete units; significant differences would then indicate that the MCAR assumption is not valid. In this case, the complete cases should not be considered a representative subsample and results will include biases. “Generally, extreme biases will occur by performing list-wise deletion when variables are MAR because the remaining datasets under-represent the population of study in question” (von Hippel, 2004 as cited in Young et al., 2011). Additionally, complete-case analysis is particularly wasteful in data sets with a large number of variables because all the incomplete cases are discarded. This can be addressed using available-case analysis.

Available-Case Analysis

Available-case analysis (or pairwise deletion) uses all the recorded data, preserving the data that is discarded in the complete-case analysis. However, this method causes the sample to change from variable to variable based on the pattern of missingness. This is particularly harmful when attempting to compare across variables, especially when the data is not MCAR. Under the assumption of MCAR, however, available-case analysis can be used to calculate means and

variances and covariances or correlations can be found using pairwise extensions which utilize all instances in which the two variables are observed. Although this pairwise method does help to recover missing data and utilize observed values, it has deficiencies of its own. These methods can yield correlations outside of the ±1 range as well as other problematic estimates that cannot be used in multivariate and contemporary analyses such as multiple regression and Structural Equation Modeling, like correlation matrices that are not positive definite (Little & Rubin, 2019; Marsh, 1998). There exists literature which found available-case analysis to be more efficient (Kim and Curry, 1977) as well as literature which supports the contrary (Haitovsky 1968; Azen et al., 1989). Little and Rubin's (2019) overall conclusion, however, is that neither method is generally satisfactory.

Schafer and Graham's (2002) simulation found similar issues with deletion methods. They found bias was produced when the assumption of MCAR was not met. They also found their results underestimated true variability. Their results tended to be higher and less variable than those values in the full population; this led to bias in both the parameter estimates and their standard errors. Literature from Pigott (2001) states that "complete-case analysis provides valid estimates under the least number of conditions and therefore is applicable to a wider range of situations than available-case analysis" (p. 365). She does acknowledge the disadvantage of being unable to anticipate if there will be enough data to remain for analysis and make results generalizable. Pigott (2001) identifies the same shortcomings of available case analysis that Little and Rubin (2019) found. She also states that there are no ways to predict when available case analysis will produce adequate results, and therefore, the method is generally not useful. Newman and Cottrell (2015) attempted to measure the bias produced by available case analysis. Their aim was to dispel the myth that this method should never be used, by finding guidelines for

how to minimize the bias and therefore make the results comparable to those obtained by maximum likelihood or multiple imputation methods. Ultimately, they found that available case analysis is preferable to complete case analysis, but "there are no known circumstances in which available case analysis will produce more accurate results than maximum likelihood or multiple imputation approaches" (p. 158).

Weighting Procedures

Building on these ideas of complete and available case analysis is another method known as weighting procedures. Weighting procedures essentially attempt to correct the biases of the deletion methods discussed above. Complete or available cases are given weights so that their distribution more closely resembles the population; this helps to produce a more randomized "sample." Weighting procedures are computationally easy to implement, and they are most useful when covariate information is limited and the sample size is large (Little & Rubin, 2019). Schafer and Graham (2002) state that weights are easy to apply for univariate and monotone missingness patterns. However, these corrections still leave something to be desired, as weighting becomes more complicated with increasingly intricate missingness patterns. These patterns can be addressed with methods that impute and model-based methods. Additionally, weighting has been found to be inefficient as weighting does not properly correct for bias and "it actually makes things worse when the adjustment cell [the weighted variable] is related to nonresponse by increasing sampling variance (Little & Rubin, 2019, p. 55).

Item Nonresponse and Single Imputation Methods

Item nonresponse is a pattern of missingness which occurs when partial data is unavailable for a given participant, such as when the participant responds to some items on a questionnaire but not all. This was traditionally handled by single imputation methods.

Imputation methods fill in the missing data; this allows researchers to then analyze the resultant data by standard methods. Little and Rubin (2019) offer this description: “Imputations should be conceptualized as draws from a predictive distribution of the missing values and require a method for creating a predictive distribution for the imputation based on the observed data” (p. 67). Mean imputation, regression imputation and Stochastic regression imputation are the most popular types of single imputation methods. Mean imputation simply replaces missing values with the mean of the observed values. Literature from van Buuren (2018) notes the issues with mean imputation are that it will underestimate the variance, disturb the relations between variables, bias almost any estimate other than the mean and bias the estimate of the mean when data are not MCAR. Obviously, the conclusion of van Buuren (2018) is that mean imputation should be avoided; Pigott (2001) goes so far as to say that “mean imputation cannot be recommended under any circumstances” (p. 366). Regression imputation uses other variables to provide better estimated imputations by building a model from the observed data and then filling in missing data with predictions from that model. This method, however, actually produces imputations that are “too good to be true.” van Buuren (2018) states that this method strengthens the relations in the data with correlations that are biased upwards and underestimation of variability. Stochastic regression imputation attempts to correct these biases by adding noise to the predictions in the form of a random draw from the residual distribution of the model constructed from observed data. This random noise does make this method suitable for imputation but some of the same limitations still apply. Newman and Cottrell (2015) identified the following problems with single imputation methods: Using complete-data N (as though data were complete) leads to underestimating standard errors (gives p-values that are too low; false positive conclusions); biased estimates, even when data are missing completely at random

(MCAR) (single imputation can create bias). One technique—stochastic regression imputation—is relatively unbiased (but standard errors are still too small). Schafer and Graham (2002) found the performance of single imputation to be better than deletion methods with small amounts of missing data, as only minor negative impacts on estimates and error measures occur. The overall conclusion is, again, that these issues can be eradicated with the use of multiple imputation and maximum likelihood methods.

Additional Missingness Patterns, Maximum Likelihood and Multiple Imputation

Other examples of missingness patterns include, but are not limited to univariate nonresponse, monotone and general or arbitrary patterns. In a univariate pattern, missingness is confined to one variable but a set of other items is completely observed. Univariate patterns can be filled in using imputation methods. Monotone patterns exist when one missing item causes subsequent items to be missing. This frequently occurs in longitudinal studies when attrition, or dropout, occurs, meaning a participant leaves the study and does not return. Attrition can be dealt with using maximum likelihood or multiple imputation methods. Arbitrary patterns exist when any set of variables may be missing any unit; these patterns usually require maximum likelihood or imputation methods.

Two modern methods are by far the most recommended methods for handling missing data, maximum likelihood estimation and multiple imputation. Both methods operate under the following assumptions: the joint distribution of the data is multivariate normal and the missingness mechanism is ignorable. Multivariate normal distribution means that all conditional distributions are normal on all other variables in the model. This allows researchers to utilize the properties of normal distributions to estimate expected values. Assuming the missingness mechanism is ignorable allows us to utilize these methods in cases besides just MCAR data.

These assumptions allow for more flexibility in handling missing data. These methods, in general, can be used for a wide range of patterns of missingness, and do so in a more powerful way than is offered by any of the procedures based on completely recorded units, weighting procedures or single imputation methods. Additionally, these methods account for the "incompleteness" of the data and adjust for biases accordingly.

Maximum likelihood (ML) estimation is a method that directly predicts the parameters and standard errors by choosing estimates for these values that maximize the probability of the observed data. This differs from multiple imputation in an important way, as multiple imputation fills in the missing values in a dataset. ML estimation bypasses this step to directly estimate the parameters and errors for the data. This method encompasses a broad class of procedures, all of which create a model that is specialized to the observed values, and therefore, there is no need to impute missing data or to delete any incomplete cases. In other words, the model is generated for a complete data set and inferences are made on the likelihood distribution under that model. The EM algorithm, proposed by Dempster et al. (1977), is an example of how the ML model is determined. Pigott (2001) offers the following explanation of how this EM algorithm works: the expectation (E) step computes an expected value for the sum of variables assuming we have the population mean and covariance. In other words, the expected value is computed based upon model parameters (likely from ML estimation) and the observed data (Finch, 2019). The maximization (M) step does the opposite, using the estimated value of the sum of the variables to compute the population mean and covariance. The E step is then repeated using the estimates of population mean and covariance from the previous M step to estimate a new expected sum of variables. Likewise, the M step repeats using the new estimated sum of variables to estimate population mean and covariance. These steps then cycle back and forth until the estimates do not

change substantially, meaning they have converged. So, we are assuming we know one piece and estimating the other.

As an overall evaluation of the method, van Buuren (2018) states that in this maximum likelihood method "the estimated parameters nicely summarize the available information under the assumed models for the complete data and the missing data." Newman and Cottrell (2015) state advantages of this method are that estimates have higher power, more accurate (but still conservative) standard error estimates, unbiased estimates and the results improve as *N* increases. Schafer and Graham (2002) found in simulations that ML estimation tends to be unbiased in large samples. They did find, however, that standard error estimates should be based on observed rather than estimated data. The literature from Pigott (2001) states that the inability to compute standard errors limits the usefulness of this method. Enders (2003) cited multiple studies (e.g., Arbuckle, 1996; Collins, Schafer, & Kam, 2001; Enders, 2001b, 2001c; Enders & Bandalos, 2001; Graham, Hofer, & MacKinnon, 1996; Wothke, 2000) which have "almost unequivocally demonstrated that Maximum Likelihood estimation is superior to ad hoc missing data techniques with respect to both bias and efficiency" (p. 322).

Multiple imputation (MI), in contrast to maximum likelihood, imputes values for the missing data values. MI extends likelihood-based methods by adding an extra step in which imputed data values are drawn. Little and Rubin (2019) state explicitly that when "missing data are unobserved values that would be meaningful for analysis if observed; in other words, a missing value hides a meaningful value" imputation should be used. According to Little and Rubin (2019), MI resolves all the problems of wastefulness, computational problems, biased variances and covariances and biased *p*-values and confidence intervals. Additionally, it is generally easier to calculate the standard errors for a wider range of parameters, and this wider

range is accomplished in MI as this method produces many plausible complete data sets. Multiple imputation is carried out in three steps. First, several conceivable complete data sets are created based on a statistical model of the data that is formed using a likelihood function and random error components (data augmentation). This is accomplished by making random draws from a posterior distribution of the missing values. This procedure helps to account for uncertainty in the imputed missing values (Galimard et al., 2016). Second, the different complete data sets are analyzed using standard procedures of the analysis of interest. Finally, this yields results that are combined into an "overall statistical analysis in which the uncertainty about the missing data is incorporated in the standard errors and significance tests" (van Ginkel et al., 2019). The multiple datasets are analyzed separately using the same procedures to obtain multiple parameter estimates. These multiple parameter estimates are combined into one by averaging (Cheema, 2014). In other words, the imputed values created by MI can be inspected and analyzed, which helps us to gauge the effect of the model assumptions on the inferences; this is not possible in maximum likelihood methods. Rubin (1987) found that the standard errors of the parameter estimates produced by this method are unbiased due to the fact that "multiple imputation specifically models the natural variation in missing data" (Cheema, 2014, p. 494).

Plenty of literature exists in favor of this method for handling missing data. Schafer and Graham (2002) recognized that the validity of MI rests on how the imputations are created and how that procedure relates to the analyses of interest but that the strength and flexibility of MI comes from the possibility of using different models for imputation and analysis. Another advantage of this method put forth by Schafer and Graham (2002) is the fact that found that "one good set of *m* imputations may effectively solve the missing data problems in many analyses; one does not necessarily need to re-impute for every new analysis" (p. 165). van Buuren (2018)

states that the goal of multiple imputation is “to obtain estimates of the scientific estimand in the population. This estimate should on average be equal to the value of the population parameter. Moreover, the associated confidence intervals and hypothesis tests should achieve at least the stated nominal value.” Although this is a high bar, plenty of evidence exists to suggest that this method can achieve these goals and that is why van Buuren (2018) explicitly states these goals. Pigott’s (2001) simulations left her “inclined to accept the results from the model-based procedures” (p. 381). van Ginkel, et. al. (2019) argued that from a theoretical point of view, MI is always to be preferred over complete case analysis, available case analysis and single imputation methods. Results from Finch (2015) even found that MI can be utilized in models that do not necessarily follow the framework of parameter estimation in his simulation in the context of multivariate analysis of variance, under many conditions. In a review by Yadav and Roychoudhury (2017) they stated that MI has better performance and multiple advantages over deletion, single imputation and maximum likelihood methods.

Although the other methods are still used in statistical analyses, most of the literature supports the use of MI or ML estimation as the optimal methods for handling missing data when the missingness mechanism is ignorable (e.g., Kang, 2013; Little & Rubin, 2019; Pigott, 2001; Schafer & Graham, 2002). Other methods may be valid in some instances and can be used accordingly. Many variations in models for MI exist and researchers need to take care in constructing effective and appropriate imputation models. One can even apply MI and ML together, facilitating MI by “utilizing EM estimates as starting values for the data augmentation algorithm” (Enders, 2010, as cited in Dong & Peng, 2013, p. 15). When utilized carefully and correctly, multiple imputation appears to be unrivaled in its ability to handle missing data values. Multiple imputation, however, does not address the problem of MNAR, as it assumes the

missingness mechanism to be MAR. Peugh and Enders (2004) state results from ML and MI "will be biased if missingness is due to the outcome variable itself (i.e., data are MNAR) (p. 535).

One variation of multiple imputation uses chained equations to address missing values in several variables. Multivariate imputation using chained equations (MICE) essentially assigns an imputation model (a regression model) for each variable with missing values and then imputes values from the conditional distribution of each variable with missing values (White et al., 2011). This process is repeated in an iterative manner for each variable with missing values. After several imputations, "the process is considered to give the correct posterior distribution of the missing data" (Galimard et al., 2016, p. 2). Galimard et al. (2016) consider MICE to be a simple, flexible and practical method to generate final imputations. MICE appears to be a powerful method for imputing missing data, particularly when data are missing on multiple variables, and is effective in its ability to handle different variable types, such as continuous, binary and categorical. White et al. (2011) state "almost any fraction of data can be validly imputed, provided that the imputation is done correctly and the MAR assumption is correct, but that any imperfections in the imputation procedure and any departures from MAR will have a proportionately larger impact when larger fractions of data are imputed" (p. 395). As this method is largely examined in the context of ignorable missingness mechanisms, sensitivity analysis is required to determine the extent of any deviation from MAR in order to understand how sensitive one's results are to potential or known MNAR mechanisms (van Buuren, 2018). Additionally, "published sensitivity analyses based on MNAR models have been largely limited to the relatively simple problem where missing values are confined to a single variable" furthering the idea that research on MNAR mechanisms is lacking and the need for methods

which can handle uncertainty in the missingness mechanism is great (Little & Rubin, 2019, p. 393).

Model-Based MNAR Methods

So, what if one does not heed Pigott's (2001) advice and investigate possible MNAR confounds during the data collection phase? If MNAR data appear in the dataset and one did not investigate the MNAR confounds, it is most likely too late to find the reason for the missing values. There are many reasons for this, such as the data collected is anonymous, participants are difficult to locate/contact after the study has ended, etc. Thus, one is left with MNAR data and without a means to recover the actual data or determine the true reason for the missing data. In this case, there exist a few MNAR methods which are discussed here. These methods are still gaining traction in social sciences research.

Enders (2011a) states that "although MAR is often reasonable, there are situations where this assumption is unlikely to hold, leading to biased parameter estimates." In his 2011(b) study, Enders demonstrates the use of three MNAR models for longitudinal data. He describes the Selection model which augments the growth curve model with a set of logistic regression equations that predict the probability of missing data at a particular wave. Like the previously mentioned methods for handling missing data, this model is accompanied by strict assumptions about normality and depends on other factors in order to model accurately and prevent substantial bias. The Shared Parameter model is similar to the Selection model, with the difference being that the individual growth curves, in this case, are used as predictors of missingness rather than the repeated measures variables. Again, this model requires multiple assumptions which are untestable and if violated can produce biased parameter estimates. The Pattern Mixture model divides the sample into subgroups based on missingness patterns and

estimates separate growth models for each subgroup. Though the assumptions are different for this model, they are present and if violated, introduce a degree of bias in model estimates.

Enders (2011a) examines a few possible options for MNAR models in latent growth curve analyses. His analyses provide "discrepant estimates," which is relatively common in such analyses. His conclusion is that "although somewhat disconcerting, it is impossible to provide general recommendations about model selection because every analytic option relies on one or more untestable assumptions. Additionally, a MAR model and an MNAR model may produce identical fit to the observed data but offer fundamentally different predictions about the unobserved score values (Molenberghs & Kenward, 2007). Enders (2011a) advises researchers to choose the model with the most defensible set of assumptions and construct a logical argument to defend one's choice. His article goes on to provide an outline for a few substantive considerations to help guide researchers in choosing a model for MNAR data. His ultimate conclusions are that "despite their limitations, these models are important options to consider, particularly when outcome-related attrition seems plausible. At the very least, MMAR models can augment the results from an MAR-based analysis."

Imputation Using Heckman's Model

The sample selection method was proposed by James J. Heckman (1976a) to address "bias that results from using nonrandomly selected samples…because of a missing data problem" (p. 153). Heckman's specific concern was the selection bias that arises through "self selection by the individuals or data units being investigated;" meaning the individual chooses to omit responses or does not take part in a study due to the data he/she would need to provide (1976b). This is exactly the idea of a MNAR mechanism in missing data where the likelihood of data being missing depends on the value of the missing data itself. This method uses two joined

linear equations, the selection equation and the outcome equation. The selection equation is obtained from a sample with no missing data (termed a selected sample) and the outcome equation is obtained from the sample with missing data in the outcome (the incomplete sample). These equations are joined by their error terms through a bivariate normal distribution (Heckman, 1976a). Galimard et al. (2016) state that Heckman's model uses a two-step estimator where the first step estimates parameters of the selection equation and the second step allows for the calculation of unbiased estimates of the outcome equation using a correction term which is obtained from the first step estimates. Of note, Heckman's model "implies the inclusion of different sets of covariates, to avoid collinearity issues" (p. 2). Therefore, a covariate that is not directly related to the outcome is included in the selection equation only; this is known as the exclusion-restriction criteria. The inclusion of this variable ideally prevents the linear predictors of the two models, selection and outcome, from being collinear (Galimard et al., 2018). The specific method for Heckman's model is outlined below.

First, the outcome equation (*Equation 1*) is defined as a linear regression equation:

$$Y_i = X_i\beta + \varepsilon_i$$

where, Y_i is a continuous outcome variable, X_i is a vector of the covariates for individual i, β is a vector of fixed effects and ε_i is the independent error term for each individual i. Next, the selection equation (*Equation 2*) is defined, representing the non-random sampling of the missingness process addressing the idea that the missingness mechanism is potentially MNAR:

$$P(R_{yi} = 1|X_i^s) = \Phi\ (X_i^s\beta^s)$$

In this selection equation: R_{yi} is a missingness indicator equal to 1 if the outcome is observed and equal to 0 if missing; X_i^s is a vector of observed variables potentially associated with

missingness; and Φ is the standard normal cumulative distribution function; β^s is an unknown vector of coefficients (Galimard et al., 2018). Essentially, this selection equation tells us the probability of Y_i being missing given the values on the vector of the observed variables. Galimard et al. state that the key point of Heckman's model is that the outcome equation ($Y_i = X_i\beta + \varepsilon_i$) and missingness indicator of the selection equation ($R^*_{yi} = X_i^s\beta^s + \varepsilon_i^s$) are linked by a bivariate normal distribution of their error terms, given a correlation coefficient, ρ, between the two error terms and a variance, $\boldsymbol{\sigma}\varepsilon^s$ of 1 (as the selection equation is a probit model with the dichotomous outcome of missing or observed). Note the vectors of observed variable values (X_i and X_i^s) and fixed effects (β and β^s) of the equations are not collinear due to the exclusion-restriction criteria (i.e., a covariate that is not directly related to the outcome is used in the selection equation and not in the outcome equation). This yields the following expectation equation (*Equation 3*) for missingness on *Y*:

$$E[R_y^*|Y] = X^s\beta^s + \rho\frac{\boldsymbol{\sigma}\varepsilon^s}{\boldsymbol{\sigma}\varepsilon}(Y - X\beta)$$

This equation shows that when $\rho = 0$, *Y* does not influence R_y and thus, the missingness mechanism is MAR. Furthermore, as ρ increases, the more *Y* has an influence R_y and the MNAR missingness mechanism becomes more important (Galimard et al., 2016, p. 4).

For each observed value, the inverse Mills Ratio (IMR) $\hat{\lambda}_i$, a control function for selection bias, is computed as:

$$\lambda_i = \frac{\phi(X_i^s\beta^s)}{\Phi(X_i^s\beta^s)}$$

where, ϕ is the standard normal density and Φ is the standard normal cumulative distribution function. That is, the IMR is the ratio of the probability density function (PDF) to the cumulative distribution function (CDF) where the PDF is used to specify the probability of the random

variable falling within a particular range of values and the CDF is the distribution function of X, evaluated at x, meaning the probability that the function will take a value less than or equal to a particular value. The IMR is then included in the following equations for the conditional mean (*Equation 4*)and variance (*Equation 5*) of *Y,* respectively:

$$E\,(Y_i \mid X_i, X_i^s, R_{yi} = 1) = X_i\beta + \rho\sigma_\varepsilon\lambda_i$$

$$Var\,(Y_i \mid X_i, X_i^s, R_{yi} = 1) = \sigma_\varepsilon^2(1 - \rho^2\delta_i) \text{ where } \delta_I = \lambda_i\,(\lambda_i + X_i^s\beta^s)$$

Given the above set of equations and conditional distributions for parameters, Heckman's procedure involves the following steps:

1) Estimate the parameters of the selection equation ($\hat{\beta}^S$) using maximum likelihood;

2) For each observed value, compute the inverse Mills Ratio (IMR) $\hat{\lambda}_i$;

3) Estimate $\hat{\beta}^S$ and a scalar coefficient with the IMR, $\hat{\beta}_\lambda$, from (*Equation 6*):

$$Y_i = X_i\beta + \hat{\lambda}_i\hat{\beta}_\lambda + \eta_i\,, \text{ where } \eta_i \text{ is the error term: } \eta \sim N\,(0, \sigma_\eta^2)$$

These steps provide unbiased estimates for the fixed effects, β. Heckman also "applied a corrector term to obtain valid estimates of variances based on a diagonal matrix of $(1 - \rho^2\delta_i)$ and adjusted to take into account for the first step estimation" as the error term is known to lack homoscedasticity (i.e., the variances are unequal across the range of predictor scores) (Heckman, 1979; Greene, 2011; Sales et al., 2004 as cited in Galimard et al., 2016, p. 4).

To end at this point would leave one with unbiased fixed effects for the original linear regression model ($Y_i = X_i\beta + \varepsilon_i$). However, to utilize Heckman's model for imputation, we must first identify the conditional expectation equation (*Equation 7*) of missing *Y:*

$$E\,(Y_i \mid X_i, X_i^s, R_{yi} = 0) = X_i\beta + \frac{-\phi(X_i^s\beta^s)}{1 - \Phi(X_i^s\beta^s)}\rho\sigma_\varepsilon$$

One can see this equation is symmetrical with the equation for the conditional mean of *Y* above, with the IMR equation substituted in and adjustments due to the equation being for

a missing *Y* as opposed to observed. This equation is used to develop the imputation model (*Equation 8*) for missing Y_i:

$$Y_i = X_i\beta + \frac{-\phi(X_i^S\beta^S)}{1-\Phi(X_i^S\beta^S)}\beta_{\lambda\ i} + \eta$$

Given the above imputation model for missing *Y* for individual *i*, the steps for imputation using Heckman's model, are:

1) Estimate all $\hat{\beta}^S$ for each predictor in *Equation 2*;

2) For each observed value, compute the inverse Mills Ratio (IMR) $\hat{\lambda}_i$;

3) Estimate $\hat{\beta}$, $\hat{\beta}_\lambda$ and $\hat{\sigma}_\eta$ using *Equation 6;*

4) Draw imputation distributions $\hat{\beta}^*$, $\hat{\beta}^*_\lambda$, $\hat{\sigma}^*_\eta$ and η^* for parameters $\hat{\beta}$, $\hat{\beta}_\lambda$, $\hat{\sigma}_\eta$ and η;

5) Impute *Y**, for each missing *Y*, using the imputation model (*Equation* 9) including distribution estimates for parameters:

$$Y^*_i = X_i\beta^* + \frac{-\phi(X_i^S\beta^S)}{1-\Phi(X_i^S\beta^S)}\beta^*_{\lambda i} + \eta^*$$

The above equation (*Equation* 9) is the imputation model for MNAR data using Heckman's model.

MICE Model for MNAR Mechanisms

Galimard et al. (2016) note that one issue of Heckman's model is it deals only with missing outcomes and subsequently handles missing covariates using complete case analysis, thus lowering power, and potentially introducing biased estimates. To simultaneously address the problem of multiple missingness mechanisms in a dataset, they propose a multiple imputation approach in the framework of chained equations for use with MNAR mechanisms that is compatible with Heckman's model. The miceMNAR model attempts to impute outcomes with missing MNAR data and predictors with missing MAR data.

Given a linear regression model, $Y \sim X_1 + X_2$, the miceMNAR procedure assigns an imputation model for each of the variables with missing data, the outcome with MNAR data, Y, and the predictor with MAR data, X_2. For the outcome, Y, the Heckman imputation model defined previously (*Equation 9*) is used with all covariates included in the outcome equation save one, which is included only in the selection equation to satisfy the exclusion-restriction criteria. The missingness indicator for Y, R_y, obtained through Heckman's model, is used in the imputation model for the predictor with MAR data, X_2, along with the same covariates used in the Heckman outcome equation for Y, yielding a linear imputation model: $X_2 \sim X_1 + Y + R_y$.

Essentially, each imputation model is a regression model which uses the other variables in the data set and an indicator of missingness in the variable. The variables with missing values, then, are sequentially imputed using the imputation models. This is repeated for several iterations and the recommended number of iterations is at least 10 (White et al., 2011; Galimard et al., 2016).

Galimard et al. (2016) generated a dataset with "three independent and identically normally distributed predictors" and a continuous predictor (p. 6). To test the ability of missing data methods, they generated 30% missing data on the outcome under three conditions: MAR and two MNAR conditions with differing levels of the importance of the MNAR mechanism (i.e., the correlation coefficient between the error terms, ρ, for Y and R_y was set to .3 or .6). They also generated 30% MAR data on one of the predictors, a simple MAR mechanism in which the missingness depended only on one other covariate. To investigate the ability of their model to impute the missing values, they first fit a linear regression model to the full dataset, prior to generating any missing values; this served as their benchmark for the subsequent analyses with missing data. Then, Galimard et al. used CCA, Heckman's selection model, standard multiple

imputation, multiple imputation using Heckman's two-step estimator and finally their own miceMNAR model (which utilized Heckman's two-step estimator in the context of MICE). For their miceMNAR model, "ten iterations of the chained equation process were considered, and the number of imputations was m = 10" (p. 7). These methods were applied to the dataset with missing values and the same linear regression was fit to all. The researchers then compared results by examining means, relative bias (the difference between the expected value of the estimate and the true value; van Buuren, 2018), root mean square of estimated standard errors, observed standard errors, root mean square error (RMSE), coverage (the proportion of confidence intervals that contain the true value; van Buuren, 2018) and percentage of the 2000 observations used by each method.

In this simulation study, Galimard et al. found, when the missingness mechanism in the outcome was known to be MNAR, the miceMNAR model provided lower RMSE, better coverage rate and less relative bias than the multiple imputation methods which rely on the assumption of a MAR mechanism. The miceMNAR model did provide larger estimated standard error values compared to multiple imputation. Overall, this method was found to be useful for data with MAR in a predictor variable and an MNAR mechanism in the outcome variable.

Current Study

Given the fact that some of these MNAR models have been slow to migrate to the social and behavioral sciences, there exists a gap in the literature for dealing with MNAR mechanisms. This study aims to provide evidence for the effectiveness of the miceMNAR model proposed by Galimard et al. (2016) and its ability to impute data and provide accurate parameter estimates. By applying multiple methods to a real dataset, we can compare the imputations and parameter estimates provided by the miceMNAR model and multivariate imputation using chained

equations (MICE). MICE was chosen for comparison due to its ability to impute data for multiple variables with missing values and the fact that it relies on the assumption that the missingness mechanism is MAR. Complete-case analysis (CCA) will also be conducted to further illustrate the importance of purposefully selecting the proper method for handling one's missing data.

MAR, considered an ignorable missingness mechanism, is not fully verifiable from the data alone when missing from observed data; a sensitivity analysis is typically required to justify the assumption of an ignorable missingness mechanism (Carpenter, et al., 2007; Galimard et al., 2016; White et al., (2011). This is because, even though one can determine a variable is related to the set of observed variables, one cannot rule out the idea that the missingness may, in some capacity, be related to the missing value itself (i.e., an MNAR mechanism). Yet, many popular methods for handling missing data rely on this assumption. Therefore, it is important to find methods which can handle the uncertainty in the missingness mechanism. The results from this study will inform researchers on potential best practices for dealing with missing data when the mechanism is unknown or known to be MNAR.

Method

Sampling and Dataset

A subset of the 2018 National Survey of Children's Health (NSCH), a survey administered to randomly selected addresses from households across the United States, was used to conduct a secondary data analysis to observe the efficacy of missing data methods. For the NSCH, one child from each household was randomly selected to be the subject of the main Topical Questionnaire; referred to as the "selected child" (Child and Adolescent Health Measurement Initiative, 2019). A sample of size $n = 2000$ was used for analyses as this matches

the value used by Galimard et al. in their 2016 simulation study. Five continuous variables were selected for use: fampov (family poverty ratio); adult1age (age, in years, of the adult identified as the primary caregiver); moved (number of times selected child has moved to a new address); HHcount (number of people living at address); childage (age of selected child, in years). These variables were selected due to their having no missing values in the dataset and the latter variables' potential for a predictive relationship with the fampov variable for use in a regression model.

Missingness Conditions - Mechanisms and Proportions of Missing Data

The miceMNAR model proposed by Galimard et al. (2016) was compared to MICE and CCA. To investigate the performance of these various methods for handling missing data, the methods were applied under varying conditions. A linear regression model was fit to the full dataset when no values were missing, serving as the benchmark for the analyses with missing data. Then, after generating missing values under the varying conditions, CCA, MICE and miceMNAR were implemented separately and the linear regression model was fit to each of the resulting datasets. Galimard et al. (2016) found their miceMNAR model able to simultaneously handle MNAR in the outcome and MAR in a predictor. Therefore, after fitting a linear regression model to the full dataset, MAR data was generated in one of the predictors for all subsequent simulations. In separate, subsequent simulations, MAR or MNAR data was generated on the outcome variable to compare the results obtained under these differing conditions. The generation of both MAR and MNAR data in different conditions allowed for an examination of the miceMNAR method's ability to handle MNAR compared to MAR. This was deemed appropriate as this method is intended to handle missing data when the mechanisms are known to be different on multiple variables. Additionally, Galimard et al. (2016) generated both MAR and

MNAR in the outcome under different conditions so, the current study aligns more closely with their methods. Varying percentages of missing data were generated for both the MAR and MNAR conditions.

Bennett (2001) found that bias is likely to occur when more than 10% of data are missing and "Enders (2003) stated that a missing rate of 15% to 20% was common in educational and psychological studies" (as cited in Dong & Peng, 2013, p. 1). Gomer and Yuan (2020), used 10%, 20%, 30% and 40% missing, Aliyara et al., (2019) used 10%, 25% and 50% missing while Enders (2003) used 15% and 30% missing for their respective studies. Galimard et al., used 30% missing data for their simulation studies in 2016 and 2018. Given this information from previous literature, 20% MAR on the predictor and 10%, 20% and 30% missing data on the outcome were selected as the missingness conditions to be utilized in this study.

Thus, analyses were conducted under with the following missingness conditions on the fampov outcome variable: a 0% missing comparison analysis, 10% MAR, 20% MAR, 30% MAR, 10% MNAR, 20% MNAR and 30% MNAR; methods and conditions for generating the missing values are explained later in this Method section. All missing data conditions also contained 20% MAR in the moved predictor. Then, CCA, MICE and miceMNAR methods were applied to the six datasets with missing values, followed by the fitting of a linear regression model to each.

Model Comparison Measures

Borrowing from the practices and terminology of Galimard et al., (2016), results were compared by obtaining the empirical mean of the parameter estimates, the relative percentage of bias, the mean of estimated standard errors and the coverage of nominal 95% confidence intervals. A comparison of the estimates obtained from the missing data techniques to the

parameter estimates from the full dataset allows for more certain determinations of which method is most effective in handling missing values under varying conditions.

Generation of Missing Data

Missing data was generated using R software, specifically the `ampute` function of the `mice` library (van Buuren & Groothuis-Oudshoorn, 2021). To illustrate, example code for the 30% MNAR in the fampov outcome (with 20%MAR in the moved predictor) condition is shown here; all subsequent discussion of `ampute` will be explained in the context of this example:

```
#generate 20% MAR in moved predictor and 30% MNAR in fampov outcome
     mypatterns<-c(1,1,0,1,1)
     mypatterns2<-rbind(mypatterns, c(1,1,1,1,0))

     myweights<-c(2, 1, 0, 1, 1)
     myweights2<-rbind(myweights, c(0,0,0,0,1))

     myfreq<-c(.40,.60)

     MNAR30<-ampute(NSCH, prop = .5, patterns = mypatterns2, freq =
     myfreq, weights = myweights2, type = "RIGHT")

     MNAR30$amp
```

The variables contained in the dataset are as follows: childage, adult1age, moved, HHcount, fampov; this ordering is relevant and will be referenced later in this section. The `ampute` function divides all cases in the dataset randomly into *k* subsets where *k* is the number of missing data patterns defined by the user. These are patterns which specify which variables will have missing values and which variables will remain complete (Schouten et al., 2018) These patterns are defined by providing a vector (if only one missingness pattern is desired) or matrix (if multiple missingness patterns are desired) for the `patterns` argument where values of 0 indicate the variable will have missing data and values of 1 indicate the variable will remain

complete. As seen above, `mypatterns` specified missingness on the continuous moved predictor which was then joined in a matrix (`mypatterns2`) with another vector specifying missingness on the continuous fampov outcome. As there were two patterns defined, all cases were assigned randomly to one of the two patterns.

The `prop` argument refers to the proportion of rows in each pattern which will have missing values. In our example, the proportion of .5 indicates that, within each pattern, 50% of the data rows will contain missing values.

The `freq` argument refers to the relative occurrence of each missingness pattern and can be specified by a vector with the length equal to the number of patterns specified. Here, we see the relative frequencies of .40 and .60. This indicates that 40% of cases were assigned to the first pattern (where the moved variable is made incomplete) and 60% of the cases were assigned to the second pattern (which generates missing data in the fampov variable). As we already specified a missingness proportion of .5, 50% of the first pattern cases have missing values, leading to 20% of the total dataset having missing values with that first pattern (.5 * .4 = .2; i.e., 20% missing in the moved variable). Similarly, 50% of the cases assigned to the second pattern have missing values, resulting in 30% missing values in the fampov variable.

Weights, and in turn the missingness mechanism of MAR or MNAR, were assigned to the missingness patterns through the `weights` argument. In essence, "the user determines which variables determine the missingness and which variables do not" through use of the weights matrix (Schouten et al., 2018). Here, the user provides a vector or matrix which specifies how the weighted sum scores for each case will be calculated. The weight matrix aligns with the pattern matrix so that the weights differ per pattern, i.e., the first vector in the weight matrix specifies weights for the cases in the first pattern vector. Weighted sum scores are the outcomes of a linear

regression model in which the coefficients are the weight values from the weight matrix. These weighted sum scores are then used to determine the probability that a case will be missing based upon additional arguments. To ampute with a MAR mechanism, variables that will have missing are given a weight of 0. To ampute with a MNAR mechanism, variables that will have missing receive a weight. Thus, to generate MAR data in the moved predictor as seen in the example above, pattern 1 was given a weight of 1 for all variables except a weight of 2 for the childage variable and a weight of 0 for the moved variable itself. The decision to give greater weight to childage was based on its use as the main control variable due to a .29 Pearson correlation between childage and moved, the highest among predictors. Additionally, there was theoretical reasoning that the propensity for missingness on the number of times a child has moved to a new address might increase with child's age as the reporting adult might be further removed from or potentially unaware of (in the cases of fostering, adopting or removal of a child/placement with relatives) any changes in a child's residential history. The weight of 0 for the moved variable specifies a MAR mechanism as the missingness will not depend on the values of the missing data. Ultimately, this means missingness was generated in the moved variable based on the scores in the observed data, a MAR mechanism, and the childage variable had a stronger relationship to whether the moved value was missing for a particular case compared to the other predictors, adult1age and HHcount. These weights were used for pattern 1 under all conditions. Weights for MNAR missingness were specified by assigning weights of 0 for all variables except the fampov outcome. This ensures the missingness depends only upon the variable in which there is missing data, a MNAR mechanism.

The `type` argument specifies the type of probability distribution which is applied to the weighted sum scores. In this case, `RIGHT` missingness was employed as this determines that

cases with high weighted sum scores will have higher probability to have missing values (Schouten et al., 2018). In the context of this study, due to the weight matrix, those with higher childage scores have a higher likelihood of having a missing value on the moved variable. Similarly, those with higher a higher family poverty ratio are more likely to have a missing value for fampov. This decision was based in theory as those with higher levels of poverty might be more likely to refrain from reporting their income.

As the weights matrix was used to create the missingness mechanisms in both patterns, the `mech` argument remained at default settings. The six missingness conditions were generated on the outcome by appropriately altering the `prop`, `freq` and `weights` arguments to adjust the percentage and missingness mechanism. Missing data was generated with 20% MAR on the moved variable for all conditions. This was a necessary step as the miceMNAR procedure was designed to handle not only MNAR data in the outcome but also MAR data in predictors.

Analysis Methods

After fitting a linear regression model to the full dataset, the following methods for handling missing data were then applied to the datasets: complete case analysis (CCA), multivariate imputation using chained equations (MICE) and the multiple imputation approach in the framework of chained equations for use with MNAR mechanisms that is compatible with Heckman's model (miceMNAR).

CCA was applied using the `na.omit` function in R which omits, or removes, all cases where any observation is missing. After dropping cases with missing values, a regression equation was fit to the CCA data.

MICE was applied using the `mice` function from the `mice` package in R (van Buuren & Groothuis-Oudshoorn, 2021). The number of imputations (argument `m`) was set to 10, yielding 10

imputed datasets of size *n* = 2000. The `method` argument was set to `NULL` and thus, was regulated by the `defaultMethod` argument, which uses Predictive Mean Matching (`pmm`) for numeric data. The `predictorMatrix` argument was left as default and, therefore, all variables in the dataset were used as predictors for each target column. The `maxit` argument which sets the number of iterations for the chained equation process, was set to 10, based on recommendations from White et al., (2011) and to match the procedure used by Galimard et al. After the cycle of imputations and iterations was complete, regression models were fit to the datasets and results were pooled.

The `miceMNAR` package was used to carry out imputation by Heckman's model for continuous outcome using a two-step estimator. First, a matrix was generated in which the selection and outcome equations are defined and joined using the `generate_JointModelEq` call with arguments `data` (where one specifies the name of dataset to be used) and `varMNAR` (where one specifies the name of the MNAR outcome to be imputed). Each of these equations is a vector of the length of the number of variables where the value is 1 if the variable is included in the equation and 0 if it is not. For example, the selection equation includes the childage, adult1age, moved and HHcount variables and therefore the vector was defined as: `c(1,1,1,1,0)`. It should also be noted that the outcome equation did not include childage, in order to meet the exclusion-restriction criteria, which requires that at least one covariate is included in the selection equation and not in the outcome equation so as to avoid collinear predictors of the two models. Childage was selected due to it having the smallest Pearson correlation (r = .04) with the fampov outcome among all predictors. The reasoning here is that childage having the smallest correlation with the outcome, not including it in the outcome

equation would have the smallest impact on the miceMNAR's ability to impute missing values while still avoiding collinearity in the two equations.

The `MNARargument` function allows the user to apply modified arguments to the `mice` function of the `mice` package for imputation of MNAR outcomes (van Buuren & Groothuis-Oudshoorn, 2021). This function simply requires the user to set the `data`, `varMNAR` and `JointModelEq` arguments as described above. The method was modified to use Heckman's two-step estimator using the following code: `arg$method["fampov"] <- "hecknorm2step"` The `mice` function was then applied to the data and arguments were specified using the `MNARargument` function's modified arguments. Iterations (`maxit`) and number of imputed datasets (`m`) were set to 10 as this matched the procedure used by Galimard et al. Example code is seen here:

```
mice.mnar.impute<- mice(data = arg$data_mod,
                        method = arg$method,
                        predictorMatrix = arg$predictorMatrix,
                        JointModelEq=arg$JointModelEq,
                        control=arg$control,
                        maxit=10,m=10)
```

Once the miceMNAR imputations were complete, regression models were fit to the datasets and results were pooled. This process was carried out on the three MAR datasets (10%, 20%, 30%) and the three MNAR datasets (10%, 20%, 30%).

Using the `for (j in 1:100)` command in R, these analyses were looped 100 times, creating 100 replications for each analysis under each missingness condition. Percent bias, the percentage of difference between the true mean and its estimate, was calculated by subtracting the full model coefficient from the simulated model coefficient for each variable across the replications for each missingness condition. 95% confidence intervals for the coefficients were obtained and used to calculate coverage probability, the proportion of times the 95% confidence

interval of the estimated summary mean contains the true value. An `ifelse` function was used to identify significant p-values for each variable across the replications for each missingness condition. This function assigns a value of 1 if the p-value is less than .05 and a value of 0 if greater than .05. Results of all analyses across the 100 replications were combined and means were calculated for coefficients, standard errors, significance, percent bias and coverage probability. These means serve as our method for comparison of the missing data methods used in this study.

Results

The mean coefficients for all methods and missingness conditions are shown in Table 1; estimates nearest the full regression are shaded for each analysis and missingness condition. Under the MAR10 and MAR20 conditions, miceMNAR provided the closest estimates for the intercept and adult1age variables and for HHcount under MAR20. Under MAR30, miceMNAR provided the closest estimates to the full regression for only HHcount. Overall, under MAR conditions, MICE and miceMNAR provided similar estimates which were close to the coefficients obtained from the full regression. Under MAR conditions, CCA performed the worst, with one exception: under MAR10 for HHcount.

MICE provided the closest estimates across the MNAR conditions with a few exceptions. miceMNAR provided the best estimate for the MNAR30 intercept. miceMNAR performed better than CCA only on the MNAR30 intercept and adult1age variable. Otherwise, miceMNAR greatly overestimated all coefficients across the MNAR conditions. It is interesting to note that CCA provided the best estimate for HHcount on three occasions: MAR10, MNAR20 (equal to MICE), MNAR30. This might indicate both MICE and miceMNAR had difficulty estimating coefficients for this particular variable. It should be noted that the moved variable contained 20%

MAR missing values across all conditions (other than the full regression model). Therefore, it is not surprising that CCA performed the worst for this variable across all MAR conditions, as both MICE and miceMNAR impute values while CCA uses only the available data.

Table 1

Mean coefficients for regression models across 100 replications by missingness condition

Coefficients						
Method parameter	MAR10	MAR20	MAR30	MNAR10	MNAR20	MNAR30
Full regression intercept	332.08	332.08	332.08	332.08	332.08	332.08
CCA intercept	324.15	318.38	306.68	326.78	317.35	307.81
MICE intercept	328.84	326.48	323.15	327.26	320.98	313.74
miceMNAR intercept	329.84	326.56	322.61	459.98	425.93	349.23
Full regression adult1age	1.46	1.46	1.46	1.46	1.46	1.46
CCA adult1age	1.54	1.70	1.91	1.33	1.19	1.11
MICE adult1age	1.53	1.59	1.63	1.41	1.36	1.32
miceMNAR adult1age	1.51	1.57	1.66	2.30	1.98	1.81
Full regression moved	-11.26	-11.26	-11.26	-11.26	-11.26	-11.26
CCA moved	-12.90	-13.55	-14.12	-12.55	-12.25	-12.05
MICE moved	-11.70	-12.12	-12.46	-11.84	-11.57	-11.56
miceMNAR moved	-12.58	-13.25	-13.71	-13.58	-13.83	-12.68
Full regression HHcount	-21.90	-21.90	-21.90	-21.90	-21.90	-21.90
CCA HHcount	-21.74	-21.33	-20.79	-22.75	-21.53	-21.30
MICE HHcount	-21.73	-21.63	-21.24	-21.69	-21.53	-21.10
miceMNAR HHcount	-22.22	-22.03	-21.87	-34.99	-31.93	-25.87

The standard errors for all analyses and missingness conditions are shown in Table 2; estimates nearest the full regression are shaded for each analysis and missingness condition. These results show that MICE provided the closest (and smallest) standard errors across all variables and missingness conditions. Very large standard errors were found for the miceMNAR

method across all conditions and variables. Especially large standard errors were obtained for all three MNAR conditions. This suggests the miceMNAR method had difficulty in estimating parameters under MNAR conditions.

Table 2

Mean standard errors for regression models across 100 iterations by missingness conditions

Standard errors						
Method parameter	MAR10	MAR20	MAR30	MNAR10	MNAR20	MNAR30
Full regression intercept	16.25	16.25	16.25	16.25	16.25	16.25
CCA intercept	20.20	21.84	24.08	19.77	21.28	22.96
MICE intercept	17.46	18.35	19.64	17.26	18.46	19.93
miceMNAR intercept	30.48	30.86	31.20	151.03	229.78	257.12
Full regression adult1age	0.29	0.29	0.29	0.29	0.29	0.29
CCA adult1age	0.36	0.40	0.44	0.36	0.39	0.42
MICE adult1age	0.31	0.32	0.34	0.30	0.32	0.35
miceMNAR adult1age	0.64	0.72	0.80	1.69	1.74	1.57
Full regression moved	1.29	1.29	1.29	1.29	1.29	1.29
CCA moved	1.70	1.91	2.22	1.60	1.72	1.87
MICE moved	1.65	1.85	2.09	1.56	1.72	1.84
miceMNAR moved	3.06	3.34	3.86	8.66	9.64	9.34
Full regression HHcount	2.45	2.45	2.45	2.45	2.45	2.45
CCA HHcount	3.13	3.43	3.82	3.03	3.26	3.51
MICE HHcount	2.65	2.87	3.11	2.61	2.77	2.96
miceMNAR HHcount	4.25	4.53	4.88	16.73	18.74	18.62

Using the `ifelse` function in R, *p*-values were assigned a 1 if below the threshold of .05 (i.e., the coefficient was significant in the regression model) and a 0 if above .05 (i.e., the coefficient was not significant in the regression model). Table 3 shows the means of these 0 or 1 values across the 100 replications for all variables and missingness conditions. This represents the percentage of times the coefficient was significant at the .05 level across 100 replications.

For example, a value of 1 indicates the coefficient was significant in every estimation of the model; a value of 0.81 indicates the coefficient was significant in 81% of the estimations of the regression model. Given that the coefficients were significant in the model based on the full dataset, a value of 1 is desirable for this parameter.

The results show that, under MAR conditions, the miceMNAR method found a significant result 100% of the time for the intercept and 97% or greater for HHcount. Results for adult1age are notable as there was significance 62%, 49%, 45% of the time for the MAR10, MAR20 and MAR30 conditions, respectively. Under the three MNAR conditions, miceMNAR found significance less than 50% of the time in all but two instances (MNAR10 intercept and MNAR10 HHcount).

Table 3

Percentage of significant coefficients across 100 iterations by missingness condition

Significance						
Method parameter	MAR10	MAR20	MAR30	MNAR10	MNAR20	MNAR30
Full regression intercept	1	1	1	1	1	1
CCA intercept	1	1	1	1	1	1
MICE intercept	1	1	1	1	1	1
miceMNAR intercept	1	1	1	0.82	0.47	0.4
Full regression adult1age	1	1	1	1	1	1
CCA adult1age	1	1	0.99	1	0.97	0.86
MICE adult1age	1	1	1	1	1	1
miceMNAR adult1age	0.62	0.49	0.45	0.35	0.20	0.18
Full regression moved	1	1	1	1	1	1
CCA moved	1	1	1	1	1	1
MICE moved	1	1	1	1	1	1
miceMNAR moved	0.94	0.93	0.89	0.47	0.38	0.45
Full regression HHcount	1	1	1	1	1	1
CCA HHcount	1	1	1	1	1	1
MICE HHcount	1	1	1	1	1	1
miceMNAR HHcount	0.97	0.97	0.99	0.68	0.44	0.34

Percent bias, the percentage of difference between the true mean and its estimate, are shown in Table 4, with the smallest bias shaded for each variable and missingness condition. It is desirable to have a bias near 0 as this indicates the estimate is very near the true mean. miceMNAR performed very well for the MAR conditions, providing the smallest, or very close to the smallest, bias for all variables and outperforming CCA in nearly all cases. miceMNAR showed a great amount of bias across the MNAR conditions, though it did provide the smallest bias for the intercept. MICE provided relatively small biases across all variables and conditions

and CCA provided the largest biases for all MAR conditions with one exception (under MAR10 on HHcount).

Table 4

Percent bias for regression coefficients across 100 iterations by missingness condition

Bias						
Method parameter	MAR10	MAR20	MAR30	MNAR10	MNAR20	MNAR30
CCA intercept	-0.02	-0.04	-0.08	-0.02	-0.04	-0.07
MICE intercept	-0.01	-0.02	-0.03	-0.01	-0.03	-0.06
miceMNAR intercept	-0.01	-0.02	-0.03	0.39	0.28	0.05
CCA adult1age	0.06	0.16	0.31	-0.08	-0.18	-0.24
MICE adult1age	0.05	0.09	0.12	-0.04	-0.06	-0.10
miceMNAR adult1age	0.04	0.08	0.14	0.58	0.36	0.24
CCA moved	0.15	0.20	0.25	0.11	0.09	0.07
MICE moved	0.04	0.08	0.11	0.05	0.03	0.03
miceMNAR moved	0.12	0.18	0.22	0.21	0.23	0.13
CCA HHcount	-0.01	-0.01	-0.10	0.08	-0.03	-0.05
MICE HHcount	-0.02	-0.02	-0.06	-0.02	-0.03	-0.07
miceMNAR HHcount	0.03	0.01	0.00	1.16	0.89	0.35

Coverage probabilities, the proportion of times the 95% confidence interval of the estimated summary mean contains the true value, are shown in Table 5. It is desirable to have coverage near 95%; coverage lower than 95% indicates an inaccurate estimator (Trikalinos et al., 2013). Values 92 - 98% are shaded in the table as these probabilities indicate good coverage. It is notable that many coverage probabilities were found to be 100% which might indicate an inefficient estimator (Trikalinos et al., 2013).

miceMNAR provided good coverage across all conditions and variables with the exception of the intercept under the MNAR10 and MNAR20 conditions. Almost all values in the

table are near 95% or greater. There are many cases in which the coverage was 100%. This is possibly an indication that these estimators were inefficient and there was perhaps not much variability across the iterations. However, knowing the large standard errors and poor coefficient estimates obtained using the miceMNAR method under MNAR mechanisms, these overall good coverage probabilities suggest the confidence intervals obtained for the miceMNAR method were quite large and therefore, quite forgiving when estimating coverage probabilities. In other words, the 95% confidence intervals were so wide, due to the inflated standard errors, that the intervals often contained the true value of the coefficient.

Table 5

Coverage probability of 95% confidence intervals across 100 iterations by missingness condition

	Coverage					
Method parameter	MAR10	MAR20	MAR30	MNAR10	MNAR20	MNAR30
CCA intercept	1	1	0.89	1	0.98	0.96
MICE intercept	1	1	1	1	1	0.98
miceMNAR intercept	1	1	1	0.89	0.89	0.94
CCA adult1age	1	0.99	0.89	1	0.99	0.98
MICE adult1age	1	1	0.99	1	1	1
miceMNAR adult1age	0.98	0.98	0.99	0.99	0.99	1
CCA moved	0.91	0.83	0.84	1	0.98	0.97
MICE moved	0.99	0.97	0.96	1	1	1
miceMNAR moved	0.99	0.97	0.96	1	1	1
CCA HHcount	1	1	1	1	1	1
MICE HHcount	1	1	1	1	1	1
miceMNAR HHcount	1	1	1	1	1	1

Discussion

Galimard et al. (2016), found their miceMNAR method to be efficient in imputing missing MNAR data in the outcome and MAR in predictors. Results from this study, however, offer little evidence in support of this method. This is not to say that the miceMNAR method is an inefficient or ineffective one; there are multiple factors which likely influenced its ability to accurately predict model parameters. The miceMNAR method resulted in highly inflated standard errors, particularly under MNAR missingness conditions. These standard error values were so large that it is difficult to trust any parameter estimates obtained from the use of this method. It is likely there were issues with the dataset and methods implemented in this study; potential issues with this study are discussed here as they relate to the poor results obtained from the miceMNAR method.

Perhaps contributing to the poor results obtained using the miceMNAR method to impute missing values is the fact that even the full dataset regression model only accounted for 9% of the variance in the outcome. This suggests these predictor variables are not accounting for very much variance and perhaps this effected the ability of the miceMNAR method to use these predictors in the joint modeling equation for imputations. It should also be noted that although all variables were continuous, values for some of the predictor variables fell within a relatively small range. For example, the moved variable ranged from 0 to 15 and the HHcount variable ranged from 2 to 15. Additionally, the childage variable used in the selection equation, ranged from 0 to 17. Although not technically discrete, these ranges might have influenced the ability of the miceMNAR method and subsequent regression models to provide accurate parameter estimates and standard errors.

MICE generally provided accurate estimates of regression coefficients and relatively small standard errors when these estimates were combined across 100 iterations of analysis. MICE provided smaller mean biases across the 100 iterations for nearly all variables and missingness conditions when compared to CCA. This supports previous research that CCA is not recommended with large amounts of missing data as CCA drops cases with missing values (Pigott, 2001). Particularly in the case of MNAR mechanisms, CCA results in biased parameter estimates as the missing value itself is related to the missingness mechanism (Ayilara et al., 2019). In this study, for example, when generating MNAR in the fampov outcome, higher levels of family poverty were more likely to be missing from the dataset. Therefore, by dropping the cases with missing values in the fampov variable (as opposed to imputing a probable value) results obtained from a regression model would underrepresent children in impoverished families, introducing bias in the resulting parameter estimates.

Although MICE assumes an ignorable missingness mechanism (MCAR or MAR), this imputation method was able to provide smaller biases in coefficient estimates across the 100 iterations of each missingness condition compared to CCA. There was one, small exception in the HHcount variable under the MAR10 condition in which biases were -0.01 and -0.02 for CCA and MICE, respectively. MICE also outperformed miceMNAR across all MNAR conditions with one, small exception in the intercept under the MNAR30 condition in which biases were 0.05 and -0.06 for miceMNAR and MICE, respectively. The miceMNAR method provided accurate coefficient estimates under MAR mechanisms and, in fact, provided biases which were smaller or very near the biases obtained from MICE. Galimard et al. (2016) found their miceMNAR model provided larger estimated standard error values compared to multiple imputation (see page 24 above) so the findings of the current study do align with their previous research in that

respect. This finding does offer support for their method in this case as the standard errors from the MAR conditions are larger but the biases are low for the miceMNAR method. However, across the MNAR conditions the standard errors are not only much larger than the standard errors from the full regression, the percent biases are quite large as well.

Further research is needed to provide evidence for the miceMNAR method in its application to real data with known nonignorable missingness mechanisms or unknown missingness mechanisms as this study failed to do so. Further research should continue to test this miceMNAR method on both simulated and non-simulated data. The variables in this dataset were selected due to their having no missing values in the full dataset. This allowed for a comparison of the missing data methods to the parameter estimates obtained from a full, non-simulated dataset. Additionally, having no missing values in the dataset allowed for a high level of control over the missingness mechanisms in the data as they were generated by the researcher. This high-level of control can be attained when using simulated data (or datasets with no missing values as was the case in this study) but the researcher then forfeits the benefit of testing their methods on real data and, ultimately, the goal of any missing data technique is to apply it to real data to preserve statistical power and obtain accurate parameter estimates from any statistical models subsequently fit to the data. In datasets with truly missing values, one can only test whether the data is MCAR or investigate possible MNAR mechanisms through sensitivity analyses, therefore adding uncertainty to the true missingness mechanism found in the data. Therefore, this miceMNAR method should be applied to both real and simulated data in future research to provide ample evidence for the efficacy of the method.

The variables, and the resulting regression model, were not necessarily selected based on any theoretical foundation and this is a limitation in this study. As mentioned previously, the full

regression model accounted for only 9% of the variance in the fampov variable, indicating a weak predictive relationship among the predictors and the outcome. This should be addressed in future research by testing the miceMNAR method on stronger statistical models. Additionally, the childage variable was selected as the exclusion-restriction criteria for the miceMNAR method due to it having a correlation of only .04 with the fampov outcome. All other variables in the outcome and selection equation were the same. Future research could investigate whether having these two equations deviate from one another even more and, in turn, further avoiding multicollinearity among these equations, might influence the imputations obtained from the joint model. Additionally, missingness could be generated on other variables in the dataset to investigate whether the reasons for the issues under MNAR mechanisms were related to something inherent in the particular variables in which missingness was generated.

This study generated 20% missing at random data in the moved variable. This percentage remained consistent across all missingness conditions generated in the outcome. Galimard et al. (2016) generated 30% missing data in the predictor and the outcome in their simulation study. Additional missingness percentages should be examined in future research. While both 20% and 30% align with Enders' (2003) finding that 15% to 20% missing data is common in psychological studies, Bennett (2001) found that bias is likely to occur when more than 10% of data are missing. Therefore, the performance of this miceMNAR method should be tested under varying amounts of missing data in the predictor as this would allow for more general statements about its ability to handle real missing data, which is likely to vary from 20% or 30%.

The results of this study failed to offer evidence for the use of miceMNAR with MAR data in predictors and MNAR data in the outcome variable. The standard errors obtained after imputation using this method were much too large for the results to be accepted with any

confidence. The limitations discussed here surely inhibited this miceMNAR method from performing optimally. With the suggested alterations to the methods put forth in this discussion, future research should be able to provide stronger support for this method.

References

Ayilara, O. F., Zhang, L., Sajobi, T. T., Sawatzky, R., Bohm, E., & Lix, L. M. (2019). Impact of missing data on bias and precision when estimating change in patient-reported outcomes from a clinical registry. *Health and Quality of Life Outcomes*, *17*(1), 106. https://doi.org/10.1186/s12955-019-1181-2

Azen, S. P., van Guilder, M., & Hill, M. A. (1989). Estimation of parameters and missing values under a regression model with non-normally distributed and non-randomly incomplete data. *Statistics in Medicine*, *8*(2), 217–228. https://doi.org/10.1002/sim.4780080208

Carpenter, J. R., Kenward, M. G., & White, I. R. (2007). Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Statistical Methods in Medical Research*, *16*(3), 259–275. https://doi.org/10.1177/0962280206075303

Cheema, J. R. (2014). A Review of Missing Data Handling Methods in Education Research. *Review of Educational Research*, *84*(4), 487–508. https://doi.org/10.3102/0034654314532697

Child and Adolescent Health Measurement Initiative (2019). "2018 National Survey of Children's Health, Sampling and Survey Administration." Data Resource Center for Child and Adolescent Health, supported by Cooperative Agreement 1-U59-MC06980-01 from the U.S. Department of Health and Human Services, Health Resources and Services Administration (HRSA), Maternal and Child Health Bureau (MCHB). Available at www.childhealthdata.org. Revised 10/7/19

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

Dong, Y., & Peng, C. Y. (2013). Principled missing data methods for researchers. *SpringerPlus*, 2(1), 222. https://doi.org/10.1186/2193-1801-2-222

Enders, C. K. (2003). Using the Expectation Maximization Algorithm to Estimate Coefficient Alpha for Scales With Item-Level Missing Data. *Psychological Methods*, *8*(3), 322–337. https://doi.org/10.1037/1082-989x.8.3.322

Enders, C. (2011a). Supplemental Material for Missing Not at Random Models for Latent Growth Curve Analyses. *Psychological Methods*, *16*(1), 1–16. https://doi.org/10.1037/a0022640.supp

Enders, C. K. (2011b). Analyzing longitudinal data with missing values. *Rehabilitation Psychology*, *56*(4), 267–288. https://doi.org/10.1037/a0025579

Finch, W. H. (2015). Missing Data and Multiple Imputation in the Context of Multivariate Analysis of Variance. *The Journal of Experimental Education*, *84*(2), 356–372. https://doi.org/10.1080/00220973.2015.1011594

Finch, W. H. (2019, May 14). *Missing Data* [PowerPoint slides]. Department of Educational Psychology, Ball State University. https://bsu.instructure.com/courses/85358/files

Galimard, J. -. E., Chevret, S., Protopopescu, C., & Resche-Rigon, M. (2016). A multiple imputation approach for MNAR mechanisms compatible with Heckman's model. *Statistics in Medicine*, *35*(17), 2907–2920. https://doi.org/10.1002/sim.6902

Galimard, J.-E., Chevret, S., Curis, E., & Resche-Rigon, M. (2018). Heckman imputation models for binary or continuous MNAR outcomes and MAR predictors. *BMC Medical Research Methodology*, *18*(1), 90. https://doi.org/10.1186/s12874-018-0547-1

Galimard, J.-E., Resche-Rigon, M. (2018). miceMNAR: Missing not at random imputation models for multiple imputation by chained equation. R package version 1.0.2. https://cran.r-project.org/web/packages/miceMNAR/miceMNAR.pdf

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

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In S. V. Berg (Eds.), *Annals of Economic and Social Measurement* (Vol. 5, Number 4, pp. 475–492). NBER.

Heitjan, D., & Basu, S. (1996) Distinguishing "Missing at Random" and "Missing Completely at Random." *The American Statistician*, 50(3), 207-213, https://doi.org/10.1080/00031305.1996.10474381

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402. https://doi.org/10.4097/kjae.2013.64.5.402

Kim, J. & Curry, J. (1977) The treatment of missing data in multivariate analysis. *Sociological Methods & Research.* 6(2), 215–40. https://doi.org/10.1177/004912417700600206.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data (Wiley Series in Probability and Statistics)* (3rd ed.). Wiley.

Marsh, H. W. (1998). Pairwise deletion for missing data in structural equation models: Nonpositive definite matrices, parameter estimates, goodness of fit, and adjusted sample sizes. *Structural Equation Modeling: A Multidisciplinary Journal*, *5*(1), 22–36. https://doi.org/10.1080/10705519809540087

Molenberghs, G., & Kenward, M. (2007). *Missing Data in Clinical Studies* (1st ed.). Wiley.

Newman, D. A. & Cottrell, J. M. (2015) Missing Data Bias: How Bad Is Pairwise Deletion? In C. E. Lance & R. J. Vandenberg (Eds.), *More Statistical and Methodological Myths and Urban Legends* (pp. 133–161). Routledge Taylor & Francis Group.

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*(4), 525–556. https://doi.org/10.3102/00346543074004525

Pigott, T. D., (2001) A review of methods for missing data. *Educational Research and Evaluation* 7(4), 353–83. https://doi.org/10.1076/edre.7.4.353.8937.

Rockel, T. (2020). missMethods: Methods for Missing Data. R package version 0.2.0. https://CRAN.R-project.org/package=missMethods

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys (Wiley Series in Probability and Statistics)* (99th ed.). Wiley.

Schafer, J. L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, *8*(1), 3–15. doi: 10.1177/096228029900800102

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

Schouten, R. M., Lugtig, P., & Vink, G. (2018). Generating missing values for simulation purposes: a multivariate amputation procedure. *Journal of Statistical Computation and Simulation*, *88*(15), 2909–2930. https://doi.org/10.1080/00949655.2018.1491577

Toomet, O., Henningsen, A., (2008). Sample Selection Models in R: Package sampleSelection. *Journal of Statistical Software.* 27(7). https://www.jstatsoft.org/v27/i07/

Trikalinos, T. A., Hoaglin, D. C., & Schmid, C. H. (2013). An empirical comparison of univariate and multivariate meta-analyses for categorical outcomes. *Statistics in Medicine*, *33*(9), 1441–1459. https://doi.org/10.1002/sim.6044

van Buuren, S. (2018). *Flexible Imputation of Missing Data, Second Edition (Chapman & Hall/CRC Interdisciplinary Statistics)* (2nd ed.). Chapman and Hall/CRC.

van Buuren S. & Groothuis-Oudshoorn K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*. 45(3), 1–67. https://www.jstatsoft.org/v45/i03/

van Buuren, S. & Groothuis-Oudshoorn, K. (2021). mice: multiple imputation by chained equations. R package version 3.13.0. https://cran.r-project.org/web/packages/mice/mice.pdf

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2019). 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

von Hippel, P. T. (2007). Regression with missing Ys: An improved strategy for analyzing multiply imputed data. Sociological Methodology, 37, 83–117. http://dx.doi.org/10.1111/j.1467-9531.2007.00180.x

White, I. R., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Yadav, M. L., & Roychoudhury, B. (2018). Handling missing values: A study of popular imputation packages in R. *Knowledge-Based Systems*, *160*, 104–118. https://doi.org/10.1016/j.knosys.2018.06.012

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. https://doi.org/10.1080/14639220903470205

Appendix A

R Code for 10% MAR condition:

```
library(foreign)
library(missMethods)
library(mice)
library(miceMNAR)
NSCH<-read.spss("/Users/brendanshanahan/Desktop/Thesis/thesis.data.sav", to.data.frame=TRUE,
use.value.labels=FALSE)

coefficients.MAR10<-NULL
ses.MAR10<-NULL
bias.MAR10<-NULL
cover.MAR10<-NULL
sigs.MAR10<-NULL


#ANALYZE COMPLETE DATA#
full.regression<-lm(fampov ~ adult1age + moved + HHcount, data = NSCH)
full.regression.sum<-summary(full.regression)
full.regression.int<-full.regression.sum$coefficients[1,1]
full.regression.adult1age<-full.regression.sum$coefficients[2,1]
full.regression.moved<-full.regression.sum$coefficients[3,1]
full.regression.hhcount<-full.regression.sum$coefficients[4,1]
full.regression.int.se<-full.regression.sum$coefficients[1,2]
full.regression.adult1age.se<-full.regression.sum$coefficients[2,2]
full.regression.moved.se<-full.regression.sum$coefficients[3,2]
full.regression.hhcount.se<-full.regression.sum$coefficients[4,2]
full.regression.int.p<-full.regression.sum$coefficients[1,4]
full.regression.adult1age.p<-full.regression.sum$coefficients[2,4]
full.regression.moved.p<-full.regression.sum$coefficients[3,4]
full.regression.hhcount.p<-full.regression.sum$coefficients[4,4]
full.regression.int.sig<-ifelse(full.regression.int.p<=0.05,1,0)
full.regression.adult1age.sig<-ifelse(full.regression.adult1age.p<=0.05,1,0)
full.regression.moved.sig<-ifelse(full.regression.moved.p<=0.05,1,0)
full.regression.hhcount.sig<-ifelse(full.regression.hhcount.p<=0.05,1,0)

#LOOPS 100 TIMES
for (j in 1:100){

#generate 20% MAR in moved predictor and 10% MAR in fampov outcome
  mypatterns<-c(1,1,0,1,1)
  mypatterns2<-rbind(mypatterns, c(1,1,1,1,0))
  myweights<-c(2, 1, 0, 1, 1)
  myweights2<-rbind(myweights, c(1,1,1,1,0))
  myfreq<-c(.67,.33)
  MAR10<-ampute(NSCH, prop = .3, patterns = mypatterns2, freq = myfreq, weights = myweights2,
type = "RIGHT")
  MAR10$amp


  #ANALYZE LISTWISE DATA#
  CCA<-na.omit(MAR10$amp)
  CCA.regression<-lm(fampov ~ adult1age + moved + HHcount, data = CCA)
  cca.regression.sum<-summary(CCA.regression)
  cca.regression.int<-cca.regression.sum$coefficients[1,1]
  cca.regression.adult1age<-cca.regression.sum$coefficients[2,1]
  cca.regression.moved<-cca.regression.sum$coefficients[3,1]
  cca.regression.hhcount<-cca.regression.sum$coefficients[4,1]
```

```
  cca.regression.int.se<-cca.regression.sum$coefficients[1,2]
  cca.regression.adult1age.se<-cca.regression.sum$coefficients[2,2]
  cca.regression.moved.se<-cca.regression.sum$coefficients[3,2]
  cca.regression.hhcount.se<-cca.regression.sum$coefficients[4,2]
  cca.regression.int.p<-cca.regression.sum$coefficients[1,4]
  cca.regression.adult1age.p<-cca.regression.sum$coefficients[2,4]
  cca.regression.moved.p<-cca.regression.sum$coefficients[3,4]
  cca.regression.hhcount.p<-cca.regression.sum$coefficients[4,4]
  cca.regression.int.sig<-ifelse(cca.regression.int.p<=0.05,1,0)
  cca.regression.adult1age.sig<-ifelse(cca.regression.adult1age.p<=0.05,1,0)
  cca.regression.moved.sig<-ifelse(cca.regression.moved.p<=0.05,1,0)
  cca.regression.hhcount.sig<-ifelse(cca.regression.hhcount.p<=0.05,1,0)

  cca.int.bias<-(cca.regression.int-full.regression.int)/full.regression.int
  cca.adult1age.bias<-(cca.regression.adult1age-
full.regression.adult1age)/full.regression.adult1age
  cca.moved.bias<-(cca.regression.moved-full.regression.moved)/full.regression.moved
  cca.hhcount.bias<-(cca.regression.hhcount-full.regression.hhcount)/full.regression.moved

  cca.ci<-confint(CCA.regression)
  cca.regression.int.cover<-ifelse(cca.ci[1,1]<full.regression.int &
cca.ci[1,2]>full.regression.int,1,0)
  cca.regression.adult1age.cover<-ifelse(cca.ci[2,1]<full.regression.adult1age &
cca.ci[2,2]>full.regression.adult1age,1,0)
  cca.regression.moved.cover<-ifelse(cca.ci[3,1]<full.regression.moved &
cca.ci[3,2]>full.regression.moved,1,0)
  cca.regression.hhcount.cover<-ifelse(cca.ci[4,1]<full.regression.hhcount &
cca.ci[4,2]>full.regression.hhcount,1,0)


  ##STANDARD MICE##
  mice.impute<-mice(MAR10$amp, print=FALSE, maxit=10, m=10)
  mice.regression<-with(mice.impute, lm(fampov ~ adult1age + moved + HHcount))
  mice.regression.pooled<-pool(mice.regression)
  mice.regression.sum<-summary(mice.regression.pooled)
  mice.regression.int<- mice.regression.sum[1,2]
  mice.regression.adult1age<- mice.regression.sum[2,2]
  mice.regression.moved<- mice.regression.sum[3,2]
  mice.regression.hhcount<- mice.regression.sum[4,2]
  mice.regression.int.se<-mice.regression.sum[1,3]
  mice.regression.adult1age.se<-mice.regression.sum[2,3]
  mice.regression.moved.se<-mice.regression.sum[3,3]
  mice.regression.hhcount.se<-mice.regression.sum[4,3]
  mice.regression.int.p<-mice.regression.sum[1,6]
  mice.regression.adult1age.p<-mice.regression.sum[2,6]
  mice.regression.moved.p<-mice.regression.sum[3,6]
  mice.regression.hhcount.p<-mice.regression.sum[4,6]
  mice.regression.int.sig<-ifelse(mice.regression.int.p<=0.05,1,0)
  mice.regression.adult1age.sig<-ifelse(mice.regression.adult1age.p<=0.05,1,0)
  mice.regression.moved.sig<-ifelse(mice.regression.moved.p<=0.05,1,0)
  mice.regression.hhcount.sig<-ifelse(mice.regression.hhcount.p<=0.05,1,0)

  mice.int.bias<-(mice.regression.int-full.regression.int)/full.regression.int
  mice.adult1age.bias<-(mice.regression.adult1age-
full.regression.adult1age)/full.regression.adult1age
  mice.moved.bias<-(mice.regression.moved-full.regression.moved)/full.regression.moved
  mice.hhcount.bias<-(mice.regression.hhcount-full.regression.hhcount)/full.regression.moved

  #confidence intervals for mids/mira
  library(parameters)
  mice.regression.pooled.parameters<-model_parameters(mice.regression.pooled, ci=0.95)
```

```
  library(dplyr)
  mice.ci<-select(mice.regression.pooled.parameters, Coefficient, CI_low, CI_high)

  #mice.ci<-confint(mice.reg.pooled)
  mice.regression.int.cover<-ifelse(mice.ci[1,2]<full.regression.int &
mice.ci[1,3]>full.regression.int,1,0)
  mice.regression.adult1age.cover<-ifelse(mice.ci[2,2]<full.regression.adult1age &
mice.ci[2,3]>full.regression.adult1age,1,0)
  mice.regression.moved.cover<-ifelse(mice.ci[3,2]<full.regression.moved &
mice.ci[3,3]>full.regression.moved,1,0)
  mice.regression.hhcount.cover<-ifelse(mice.ci[4,2]<full.regression.hhcount &
mice.ci[4,3]>full.regression.hhcount,1,0)


  ###SELECTION MODEL WITH MICE###

  # Import dataset with a suspected MNAR mechanism
  require(GJRM)
  library(miceMNAR)
  # Specify a selection (missing data mechanism) and an outcome equation (analyse model)
  # Generate an empty matrix
  JointModelEq <- generate_JointModelEq(data=MAR10$amp,varMNAR = "fampov")
  # Fill in with 1 for variable included in equations
  JointModelEq[,"fampov_var_sel"] <- c(1,1,1,1,0)
  # This indicates that childage, adult1age, moved and HHcount are included in the selection
equation of fampov
  JointModelEq[,"fampov_var_out"] <- c(0,1,1,1,0)
  # This indicates that adult1age, moved and HHcount are included in the outcome equation of
fampov

  ## Using 2-step estimation ##
  arg <- MNARargument(data=MAR10$amp,varMNAR="fampov",JointModelEq=JointModelEq)
  arg$method["fampov"] <- "hecknorm2step"

  mice.mnar.impute<- mice(data = arg$data_mod,
                          method = arg$method,
                          predictorMatrix = arg$predictorMatrix,
                          JointModelEq=arg$JointModelEq,
                          control=arg$control,
                          maxit=10,m=10)
  mice.mnar.regression<-with(mice.mnar.impute, lm(fampov ~ adult1age + moved + HHcount))
  mice.mnar.regression.pooled<-pool(mice.mnar.regression)
  mice.mnar.regression.sum<-summary(mice.mnar.regression.pooled)
  mice.mnar.regression.int<- mice.mnar.regression.sum[1,2]
  mice.mnar.regression.adult1age<- mice.mnar.regression.sum[2,2]
  mice.mnar.regression.moved<- mice.mnar.regression.sum[3,2]
  mice.mnar.regression.hhcount<- mice.mnar.regression.sum[4,2]
  mice.mnar.regression.int.se<-mice.mnar.regression.sum[1,3]
  mice.mnar.regression.adult1age.se<-mice.mnar.regression.sum[2,3]
  mice.mnar.regression.moved.se<-mice.mnar.regression.sum[3,3]
  mice.mnar.regression.hhcount.se<-mice.mnar.regression.sum[4,3]
  mice.mnar.regression.int.p<-mice.mnar.regression.sum[1,6]
  mice.mnar.regression.adult1age.p<-mice.mnar.regression.sum[2,6]
  mice.mnar.regression.moved.p<-mice.mnar.regression.sum[3,6]
  mice.mnar.regression.hhcount.p<-mice.mnar.regression.sum[4,6]
  mice.mnar.regression.int.sig<-ifelse(mice.mnar.regression.int.p<=0.05,1,0)
  mice.mnar.regression.adult1age.sig<-ifelse(mice.mnar.regression.adult1age.p<=0.05,1,0)
  mice.mnar.regression.moved.sig<-ifelse(mice.mnar.regression.moved.p<=0.05,1,0)
  mice.mnar.regression.hhcount.sig<-ifelse(mice.mnar.regression.hhcount.p<=0.05,1,0)

  mice.mnar.int.bias<-(mice.mnar.regression.int-full.regression.int)/full.regression.int
```

```
  mice.mnar.adult1age.bias<-(mice.mnar.regression.adult1age-
full.regression.adult1age)/full.regression.adult1age
  mice.mnar.moved.bias<-(mice.mnar.regression.moved-
full.regression.moved)/full.regression.moved
  mice.mnar.hhcount.bias<-(mice.mnar.regression.hhcount-
full.regression.hhcount)/full.regression.moved

  #confidence intervals for mids/mira
  library(parameters)
  mice.mnar.regression.pooled.parameters<-model_parameters(mice.mnar.regression.pooled,
ci=0.95)
  library(dplyr)
  mice.mnar.ci<-select(mice.mnar.regression.pooled.parameters, Coefficient, CI_low, CI_high)

  mice.mnar.regression.int.cover<-ifelse(mice.mnar.ci[1,2]<full.regression.int &
mice.ci[1,3]>full.regression.int,1,0)
  mice.mnar.regression.adult1age.cover<-ifelse(mice.mnar.ci[2,2]<full.regression.adult1age &
mice.ci[2,3]>full.regression.adult1age,1,0)
  mice.mnar.regression.moved.cover<-ifelse(mice.ci[3,2]<full.regression.moved &
mice.ci[3,3]>full.regression.moved,1,0)
  mice.mnar.regression.hhcount.cover<-ifelse(mice.ci[4,2]<full.regression.hhcount &
mice.ci[4,3]>full.regression.hhcount,1,0)


  #COMBINE ESTIMATES#
  coefficients<-cbind(full.regression.int, full.regression.adult1age, full.regression.moved,
full.regression.hhcount,
                      cca.regression.int, cca.regression.adult1age, cca.regression.moved,
cca.regression.hhcount, mice.regression.int, mice.regression.adult1age,
                      mice.regression.moved, mice.regression.hhcount,
mice.mnar.regression.int, mice.mnar.regression.adult1age, mice.mnar.regression.moved,
                      mice.mnar.regression.hhcount)

  #COMBINE STANDARD ERRORS#
  ses<-cbind(full.regression.int.se, full.regression.adult1age.se, full.regression.moved.se,
full.regression.hhcount.se,
             cca.regression.int.se, cca.regression.adult1age.se, cca.regression.moved.se,
cca.regression.hhcount.se, mice.regression.int.se,
             mice.regression.adult1age.se, mice.regression.moved.se,
mice.regression.hhcount.se, mice.mnar.regression.int.se,
             mice.mnar.regression.adult1age.se, mice.mnar.regression.moved.se,
mice.mnar.regression.hhcount.se)

  #COMBINE BIAS RESULTS#
  bias<-cbind(cca.int.bias, cca.adult1age.bias, cca.moved.bias, cca.hhcount.bias,
mice.int.bias, mice.adult1age.bias, mice.moved.bias,
              mice.hhcount.bias, mice.mnar.int.bias, mice.mnar.adult1age.bias,
mice.mnar.moved.bias, mice.mnar.hhcount.bias)


  #COMBINE COVERAGE RATES#
  cover<-
cbind(cca.regression.int.cover,cca.regression.adult1age.cover,cca.regression.moved.cover,cca.r
egression.hhcount.cover,
               mice.regression.int.cover, mice.regression.adult1age.cover,
mice.regression.moved.cover, mice.regression.hhcount.cover,
               mice.mnar.regression.int.cover, mice.mnar.regression.adult1age.cover,
mice.mnar.regression.moved.cover,
               mice.mnar.regression.hhcount.cover)

  #COMBINE SIGNIFICANCE TEST RESULTS#
```

```
  sigs<-cbind(full.regression.int.sig, full.regression.adult1age.sig,
full.regression.moved.sig, full.regression.hhcount.sig,
              cca.regression.int.sig, cca.regression.adult1age.sig, cca.regression.moved.sig,
cca.regression.hhcount.sig, mice.regression.int.sig,
              mice.regression.adult1age.sig, mice.regression.moved.sig,
mice.regression.hhcount.sig, mice.mnar.regression.int.sig,
              mice.mnar.regression.adult1age.sig, mice.mnar.regression.moved.sig,
mice.mnar.regression.hhcount.sig)

  #COMBINE RESULTS ACROSS REPLICATIONS#
  coefficients.MAR10<-rbind(coefficients.MAR10, coefficients)
  ses.MAR10<-rbind(ses.MAR10, ses)
  sigs.MAR10<-rbind(sigs.MAR10, sigs)
  bias.MAR10<-rbind(bias.MAR10, bias)
  cover.MAR10<-rbind(cover.MAR10, cover)

}

#CALCULATE MEANS FOR DIFFERENT OUTCOMES AND SAVE TO A FILE#
coefficients.mean.MAR10<-t(colMeans(coefficients.MAR10))
ses.mean.MAR10<-t(colMeans(ses.MAR10))
bias.mean.MAR10<-t(colMeans(bias.MAR10))
cover.mean.MAR10<-t(colMeans(cover.MAR10))
sigs.mean.MAR10<-t(colMeans(sigs.MAR10))

coefficients.mean.MAR10
ses.mean.MAR10
bias.mean.MAR10
cover.mean.MAR10
sigs.mean.MAR10

#WRITE RESULTS TO THE HARD DRIVE#
##YOU WILL NEED TO CHANGE THESE ADDRESSES TO MATCH WHERE YOU WANT THESE FILES STORED ON YOUR
COMPUTER##
write.table(coefficients.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/coefficients_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(ses.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/ses_resultsMAR10.out", col.names=TRUE, row.names=TRUE)
write.table(sigs.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/sigs_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(bias.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/bias_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(cover.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/cover_resultsMAR10.out", col.names=FALSE, row.names=FALSE)

write.table(coefficients.mean.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/coefficients_mean_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(ses.mean.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/ses_mean_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(sigs.mean.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/sigs_mean_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(bias.mean.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/bias_mean_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
write.table(cover.mean.MAR10, file="/Users/brendanshanahan/Desktop/Thesis/simulation
output/cover_mean_resultsMAR10.out", col.names=FALSE, row.names=FALSE)
```