Getting tough on missing data: a boot camp for social science researchers

Sinan Gemici

Alice Bednarz

Patrick Lim

NATIONAL CENTRE FOR
VOCATIONAL EDUCATION RESEARCH

NATIONAL CENTRE FOR
VOCATIONAL EDUCATION RESEARCH
TECHNICAL PAPER

NCVER

Getting tough on missing data: a boot camp for social science researchers

Sinan Gemici
Alice Bednarz
Patrick Lim

National Centre for Vocational Education Research

FOR CLARITY WITH FIGURES PLEASE PRINT IN COLOUR.

NATIONAL CENTRE FOR VOCATIONAL EDUCATION RESEARCH
TECHNICAL PAPER

The views and opinions expressed in this document are those of the author/ project team and do not necessarily reflect the views of the Australian Government, state and territory governments or NCVER.

This document should be attributed as Gemici, S, Bednarz, A & Lim, P 2011, *Getting tough on missing data: a boot camp for social science researchers*, NCVER.

The National Centre for Vocational Education Research (NCVER) is an independent body responsible for collecting, managing and analysing, evaluating and communicating research and statistics about vocational education and training (VET).

NCVER's inhouse research and evaluation program undertakes projects which are strategic to the VET sector. These projects are developed and conducted by the NCVER's research staff and are funded by NCVER. This research aims to improve policy and practice in the VET sector.

ISBN 978 1 921955 53 2 web edition
978 1 921955 54 9 print edition
TD/TNC 104.43

Published by NCVER, ABN 87 007 967 311

Level 11, 33 King William Street, Adelaide SA 5000
PO Box 8288 Station Arcade, Adelaide SA 5000, Australia

P +61 8 8230 8400 **F** +61 8 8212 3436 **E** ncver@ncver.edu.au **W** http://www.ncver.edu.au

About the research

Getting tough on missing data: a boot camp for social science researchers

Sinan Gemici, Alice Bednarz & Patrick Lim, NCVER

Research in the social sciences is routinely affected by missing or incomplete information. Ignoring missing data may yield research findings that are either 'slightly off' or 'plain wrong'. However, there is often confusion over how best to handle missing data.

In this paper, the authors repackage the highly technical missing data literature into a more accessible format. They illustrate why and how simple approaches to handling missing data fail. Here, simple approaches are those that delete records with missing data or which replace missing observations with crude estimates of their hypothesised 'true' value. They then discuss several common methods for addressing missing data and conduct a simulation study with real-life data to assess the performance of these methods. They conclude with a step-by-step guide on how to implement multiple imputation as one of two current 'gold standard' missing data methods.

The key message of this paper is that modern software packages make it relatively easy to implement methods that handle missing values properly.

Tom Karmel
Managing Director, NCVER

Contents

Appendices

Tables and figures

Tables

Figures

Introduction

Research in the social sciences is routinely affected by missing or incomplete information. Given that this is the norm rather than the exception, why are we so annoyed by missing data? Is it because missing data stand between us and the research questions we want to answer, or because missing information reduces the confidence we have in our findings? Or is it because we are simply unsure about how to handle missing data appropriately? Whatever the reason, as producers of applied social science research, we perceive missing data to be a nuisance that greatly complicates our work.

Many situations can cause information to be missing, including an undue response burden from lengthy questionnaires, erroneous data entry, or a refusal to answer questions that are considered intrusive. Missing information is problematic because most statistical procedures used in the social sciences require complete data. Consequently, ignoring incomplete information may yield research findings that are either 'slightly off' or 'plain wrong'.

Much has been written about the pitfalls associated with missing data, as well as the different methods available to address them. So why are we writing this paper? We are motivated by the fact that much contemporary social science research continues to ignore missing data problems despite the widespread availability of methods that adequately address the issue. One likely reason for disregarding these methods is that most existing literature on missing data is highly technical and presumes extensive statistical and mathematical expertise on the part of the reader. Not all social scientists have undergone sufficient training in mathematical statistics to engage with this literature. Many applied researchers also have a limited desire to grapple with the theoretical underpinnings of missing data and are much more interested in practical guidance on handling concrete missing data problems. Therefore, our objective is to repackage the highly technical missing data literature into a more accessible format. We do so specifically with those researchers in mind who analyse large-scale surveys or administrative data collections.

We commence this paper with a brief overview of the mechanisms that drive missing data and the patterns in which they occur. We then discuss several common methods for addressing missing data problems. We conclude with a simulation study in which we use data from the Longitudinal Surveys of Australian Youth (LSAY) and the National Vocational Education and Training Provider Collection to assess the performance of selected missing data methods. Practical guidelines for tackling concrete missing data problems are provided in the appendix.

We find that addressing missing data with basic methods can severely impact on research results. While using more principled methods requires additional time and effort, we argue that not doing so can mean missing the mark when generating research findings. Our paper takes researchers on a 'boot camp', which will hopefully help them to get tough on missing data problems in their own work.

Missing data 'behind the scenes'

Tackling a concrete missing data problem requires that we understand *why* and *how* information is missing. This understanding is key to identifying the *mechanism* that drives a missing data problem, which has direct implications for the particular *method* we may choose to address it. Here, we provide a concise, non-technical overview of these 'behind the scenes' aspects of missing data. Comprehensive technical overviews of missing data theory are available in the academic literature (for example, Little & Rubin 2002; Schafer 1997).

The 'why' of missing data

Data can be missing for numerous reasons. Rather than list all possible causes and scenarios, we focus on those that commonly affect large-scale observational data from surveys or administrative collections in the social sciences.

Understanding why data are missing allows us to determine the type of non-response we are dealing with. We generally differentiate between attrition, unit non-response, item non-response, and wave non-response:

- *Attrition* occurs in longitudinal surveys and means that respondents who participate in the initial survey wave drop out at some future wave. Dropout may occur for logistical reasons or because an eligible respondent loses interest in the survey. Attrition implies that the respondent does not rejoin the survey after dropping out.
- *Wave non-response* also occurs in longitudinal surveys where participants may be missing for one or more survey waves. Wave non-response is different from attrition because the former implies that respondents rejoin the survey after missing one or more waves.
- *Unit non-response* means that we have no data at all for an eligible respondent. This occurs if an eligible respondent refuses to participate in a study or survey, or if for some reason a respondent's record is lost.
- *Item non-response* refers to situations in which a respondent answers some items but fails to answer others. Item non-response frequently occurs with questions that the respondent perceives as intrusive (for example, questions about income, drug use, or sexual practices). It also occurs as a consequence of survey fatigue where lengthy and complicated questionnaires demotivate respondents. Missingness from undue response burden is exacerbated if respondents have limited interest in the survey topic or consider it irrelevant to their personal circumstances.

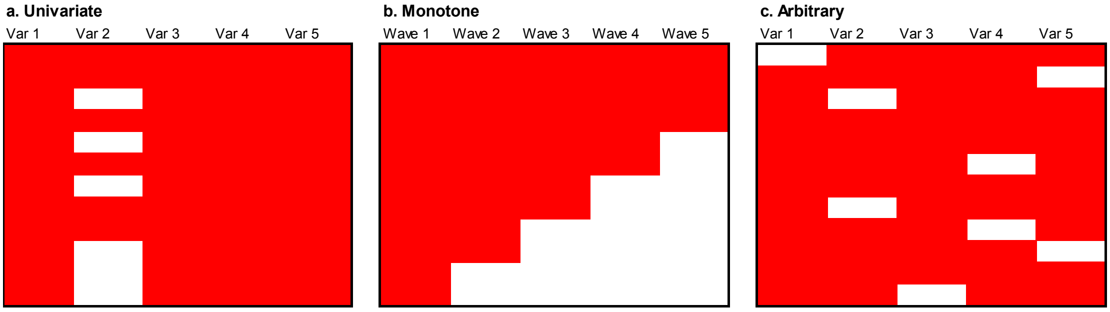
Determining the type of non-response is important because different scenarios require different missing data methods. *We emphasise that the methods discussed in this paper focus on missing data from item non-response*. Missing data for other non-response types are typically addressed using weighting methods that are illustrated elsewhere (see Lim 2011 for weighting in the Longitudinal Surveys of Australian Youth; Piesse & Kalton 2009 for wave non-response weighting).

The 'how' of missing data

When thinking about *how* data are missing we refer to missing data *patterns*. Patterns reflect the way in which missing values structurally appear in our dataset. We generally differentiate between *univariate*, *monotone*, and *arbitrary* patterns, although variations of these fundamental patterns exist.

A *univariate* pattern occurs when a specific variable contains missing values, while all other variables are fully observed (figure 1a). This can happen when, for example, respondents perceive a specific questionnaire item as particularly sensitive. *Monotone* missingness is typical of attrition in longitudinal surveys when individuals decide to drop out before all survey waves have been completed. Monotone missingness yields a staircase-like pattern with steadily increasing amounts of missing values for each data collection wave (figure 1b). Finally, patterns are *arbitrary* when the missing values make up no systematic, discernable structure within a dataset (figure 1c). This can occur when missing values are distributed randomly throughout the dataset as a result of item non-response or errors in data entry.

Figure 1 Graphical representation of missing data patterns



Note: Missing values are represented by a blank rectangle; complete values are represented by a coloured rectangle.

Alternatively, missing data patterns can be depicted in tabular form (see table 1).

Table 1 Tabular representation of missing data patterns

Group	Freq	Var1	Var2	Var3	Var4	Missing Vars
1	2475	X	X	X	X	0
2	482	X	X	X	.	1
3	319	.	X	X	X	1
4	344	X	X	.	X	1
5	308	X	.	X	X	1
6	88	.	X	X	.	2
7	161	X	X	.	.	2
8	92	.	X	.	X	2
9	75	X	.	X	.	2
10	145	.	.	X	X	2
11	141	X	.	.	X	2
12	53	.	X	.	.	3
13	66	.	.	X	.	3
14	80	X	.	.	.	3
15	87	.	.	.	X	3
16	84	.	.	.	.	4
		934	986	1042	1089	4051

Note: Missing values are depicted as '.'; observed values are depicted as 'X'.

In table 1, patterns of missingness are categorised by groups. The first group consists of 2475 records that are complete on all four variables in the dataset. In the second group, 482 records contain missing values on the fourth variable (that is, univariate missingness). In the 16th group, 84 records have missing values on all four variables. The bottom row shows the total number of missing values per variable. The total number of missing values throughout the entire dataset amounts to 4051, and most of them (1089) occur on the fourth variable. The final column summarises the number of variables with missing values per group.

The reason we are interested in the 'how' of missing values is that it can be related to the 'why'. This relationship is more obvious in fairly simple scenarios, as with the relationship between attrition and monotone patterns, or between erroneous data entry and arbitrary patterns. In many real-life situations, however, researchers may find the connections between causes and patterns to be obscured by a mixture of different missing data patterns within the same dataset. This is particularly true when working with large-scale observational or administrative data collections that feature numerous survey waves and/or variables.

Missing data mechanisms

Clarity about why and how values are missing is important in order to understand better the mechanism that drives a concrete missing data problem. Likewise, understanding this mechanism is crucial for selecting an appropriate missing data method. Missing data mechanisms capture the probabilistic relationships between missing and observed values in a dataset. The following sections further clarify this somewhat complex concept.

We distinguish between three mechanisms that can underlie a missing data problem, including *missing completely at random* (MCAR), *missing at random* (MAR), and *missing not at random* (MNAR). Introduced by Rubin (1976), this standard terminology is not directly intuitive to many researchers who begin to grapple with concrete missing data problems. However, gaining familiarity with these terms and concepts is crucial because they are used constantly in all discussions relating to missing data and these are referenced heavily throughout the remainder of this paper. Given their fundamental importance in understanding missing data, we attempt to describe these mechanisms conceptually rather than statistically. Interested readers are referred to Little and Rubin (2002) for a comprehensive technical treatise of the topic.

Missing completely at random (MCAR)

When data are *missing completely at random,* the missing observations simply represent a random sample from within all observations in the dataset. As mentioned earlier, random error in data entry represents one of many possible scenarios of completely random missingness. Since no structural association exists between missing and observed data, missing values do not alter the original distributional relationships between variables.

Missing completely at random is considered 'ignorable' because it is not necessary to model the missing data mechanism separately from modeling the parameter estimates we are interested in. In terms of missing data bias, no particular method is required to address the problem. However, discarding cases with missing data reduces sample size and statistical power. We revisit this point in more detail in our discussion of missing data methods.

We strongly caution readers that the assumption of data *missing completely at random* is unrealistic in most social science collections. This is particularly true for large-scale observational surveys that contain numerous interrelated variables. Researchers can test whether data are *missing completely at random* by conducting Little's (1988) MCAR test. This test compares sub-groups with missing and observed data that share the same missing data pattern. The null hypothesis of *data missing completely at random* is rejected if the missing values do not represent a true random sample of the observed data. Little's MCAR test is available in several standard statistical software packages and should be conducted as part of exploratory data analysis. (Additional information on Little's MCAR test is provided in appendix A: Practical guidelines for applied researchers.)

Missing at random (MAR)

While understanding the *missing completely at random* mechanism is important, very few social science datasets contain observations that are actually missing *completely* at random. A weaker condition, labelled *missing at random*, is therefore more realistic in practice. When data are *missing at random*, missing observations in a particular variable are related to one or more of the other variables in the dataset, given that these other variables are fully observed. However, missing observations must not depend on the variable in which they occur. (This scenario is called *missing not at random* and is described in the following sub-section.)

To illustrate the *missing at random* mechanism, we consider a cohort of school completers who are asked to report their tertiary entrance rank (TER) scores. We further assume that students who have been in Australia for less than five years, as well as students from lower socioeconomic backgrounds, are less likely to report their TER scores. As long as the missingness in TER scores is conditional on a student's length of in-country residence and socioeconomic profile, but *not* on the TER score itself, the assumption of the data being *missing at random* is satisfied.

The above example clarifies that, despite the term 'missing at random', missing data under this mechanism do not constitute an independent random subset of the observed data. However, *missing at random* is still considered 'ignorable' because there is no need to model the missing data mechanism separately from estimating the parameter estimates we are interested in (Allison 2002). This modelling of the missing data mechanism is required only when data are *missing not at random*, as outlined in the following sub-section.

Missing not at random (MNAR)

Missing not at random refers to situations in which, even after controlling for other variables in the dataset, the missingness on a variable is still related to the missing values in that *same* variable. To clarify, let us consider the same TER score scenario as above. Under the assumption that data are *missing not at random*, the probability of missing TER scores may or may not depend on the respondents' length of in-country residence and socioeconomic background, as well as any other variable in the dataset. What is important is that the probability would also depend on the *value of the TER score itself*. Supposing that respondents with lower TER scores are less likely to report them, the value of the TER score thus influences the probability of its missingness.

Missing not at random is considered a 'non-ignorable' mechanism because the missingness depends, at least partially, on the missing data themselves. Since we cannot verify the value of a missing observation, we have to model the missing data mechanism separately from modelling the parameter estimates we are interested in. Allison (2002) cautions that 'for effective estimation with non-ignorable missing data, very good prior knowledge about the nature of the missing data process usually is needed, because the data contain no information about what models would be appropriate and the results typically will be very sensitive to the choice of model' (p. 5). Overall, non-ignorable missingness greatly complicates the handling of incomplete data.

How can we know whether data are *missing at random* (ignorable) or *missing not at random* (non-ignorable)? The simple answer is that we cannot. Little's MCAR test only determines whether data are *missing completely at random*. Once the null hypothesis of completely random missingness is rejected, no further statistical tests are available to determine the remaining mechanisms. While Bayesian modelling techniques are available to address non-ignorable missingess (see Mason et al. 2010), we caution readers that understanding and implementing these techniques requires considerable technical expertise. A more practical option for less experienced researchers can therefore be to assume that data are *missing at random* for practical purposes and to be explicit about the potential bias in research outcomes.

Key points to remember

The take-home message from our 'behind the scenes' discussion is that we need to familiarise ourselves with the concrete missing data scenario we are faced with. This entails a consideration of reasons, patterns, and mechanisms. The latter is critical, but can be particularly challenging.

It often helps to think about missing data mechanisms in terms of their biasing effects on research results. The *missing completely at random* mechanism has no systematic impact on bias and is the easiest scenario to address. It is also the least realistic one, especially when working with large-scale social science data. A *missing at random* mechanism will likely bias our results unless we use adequate methods to address our missing data problem. A *missing not at random* mechanism can have a strong impact on bias unless it is addressed using very specific models. However, implementing and assessing these models is very costly in terms of the time and expertise required.

To help retain key terminology for missing data mechanisms, we summarise standard acronyms in table 2.

Table 2 Summary of standard acronyms for missing data mechanisms

Mechanism	Meaning	Description
MCAR	Missing completely at random	Missingness is unconditional on any specific variable in the dataset
MAR	Missing at random	Missingness is conditional on observed variables, but not on the variable on which it occurs
MNAR	Missing not at random	Missingness is conditional on the variable on which it occurs, as well as observed variables

Missing data methods

Numerous methods have been developed over time to address missing data problems. Methods range from basic case deletion to more complex imputation procedures. Rather than detail each available method, we discuss a select few that are commonly used in applied social science research.

Case deletion methods

Case deletion methods, such as listwise or pairwise deletion, are popular because they allow researchers to dispense with missing data problems quickly and easily. However, the limitations inherent in these basic methods far outweigh their benefits in many research scenarios, and especially those that involve the analysis of large-scale data.

Listwise deletion

Listwise deletion (also referred to as *complete case analysis*) is the most prominent case deletion method in social science research. Any record with a missing value on one or more variables is discarded from statistical analysis. This is done either by physically or logically removing the record from the dataset. The latter is the default setting for handling missing values in standard statistical software packages because commonly used statistical procedures have been developed for complete data.

Listwise deletion has two major limitations. First, it requires data to be *missing completely at random*. This assumption may be viable when technical issues or errors in data entry produce missing values in administrative data collections. However, missing values in large-scale surveys are most often related to participant characteristics, such as socioeconomic status, academic achievement, disability status, immigrant status, and many others. As discussed earlier, the practical implication of *missing at random* or *missing not at random* mechanisms is that respondents with observed data are systematically different from those with missing data. Listwise deletion can thus bias results from data analysis because the remaining sample is no longer representative of the original population of interest.[1]

A second important disadvantage is that listwise deletion reduces available sample size. The smaller the remaining complete-data sample, the greater the loss of statistical power, which curtails our ability to detect a significant effect through statistical testing. Listwise deletion therefore increases our risk of failing to detect potentially important relationships between predictor variables and outcomes of interest. When used with large-scale social science datasets, listwise deletion often results in particularly dramatic case loss due to the large number of variables on which missing values can occur.

Pairwise deletion

Pairwise deletion (also referred to as *complete variables analysis*) is a variable-by-variable method where only those cases that exhibit missing values on a particular bivariate pair are discarded. This approach can be used with common statistical procedures, such as correlation analysis, ANOVA, and regression. Pairwise deletion shares the disadvantages associated with listwise deletion, such as the

[1] Listwise deletion can yield unbiased parameter estimates in linear regression models even when data are *missing at random* or *missing not at random* if the missing values depend on a single predictor rather than the outcome variable (see Little 1992 for more information).

need for data to be *missing completely at random* and loss of statistical power (albeit to a lesser extent). Moreover, pairwise deletion produces biased standard errors (that is, sampling fluctuation around an estimate) due to the fact that sample sizes vary according to which bivariate pair is considered. (See Enders 2010 for a technical discussion of standard error bias in pairwise deletion.)

Single imputation methods

Whereas case deletion methods simply discard missing data, single imputation methods replace any missing data point with a simple fixed estimate of the hypothesised 'true' value. Numerous single imputation methods have been developed that vary in complexity. Here, we discuss two of the less complex single imputation methods routinely used in the social sciences.

Constant replacement

Constant replacement methods replace a missing data point with a simple fixed estimate of the unobserved value. This estimate is usually the mean or mode of the variable in which the missing data point occurs. Besides being easy to implement, the key advantage of constant replacement over case deletion lies in the preservation of sample size and, thus, statistical power. However, replacing missing values with a constant can severely reduce the variability in the data. Reduced variability leads to biased estimates of variances and covariances unless data are *missing completely at random*.

Regression imputation

Regression imputation replaces missing data with the predicted values from a linear regression model. This method requires at least a moderate degree of covariance between variables with missing data and all other variables within the data matrix. Since imputed values fall directly on the regression plane, the residual variability in the data is diminished. To offset this effect, a random error term can be added to the imputation model to introduce additional variance. This procedure is commonly referred to as *stochastic* regression imputation. Although easy to implement, regression imputation is known to overestimate correlations and R^2 values, and underestimate standard errors (Enders 2010).

Maximum likelihood estimation

Maximum likelihood estimation is a method for estimating unknown population parameters, such as the means, variances, and covariances. Given a complete sample from a population of interest, the procedure uses a likelihood function to estimate those population parameters that are *most likely* to have produced that particular sample (Enders 2010). The likelihood function is different for each sample from the population.

When our sample contains missing values, maximum likelihood estimation is more difficult, because in addition to the unknown *parameters*, we now have unknown *data* to deal with. The two methods used to perform maximum likelihood estimation on incomplete data include expectation-maximisation and direct maximum likelihood.

Expectation-maximisation

The idea behind the expectation-maximisation algorithm is to first 'fill in' the missing values and then find the maximum likelihood estimates for the complete-data problem. This is much easier than trying to directly generate maximum likelihood estimates for the incomplete-data problem (Schafer & Graham 2002). The expectation-maximisation algorithm iterates between an 'expectation' and a ‘maximisation’ step. In the expectation step, missing values are ‘filled in’ with regression and covariance estimates of the observed data, whereas the subsequent maximisation step *maximises* the likelihood function. Maximising a likelihood function means finding those estimates of the population parameters that are most likely (that is, have the ‘*maximum likelihood*’) to match/produce the sample of data we are working with. The algorithm iterates between these two steps using continuously updated estimates, meaning that the maximisation step recalculates parameters based on the re-estimated filled-in missing data parameters from the expectation step until the parameter estimates from both steps converge (that is, they hardly change from one iteration to the next).

While expectation-maximisation can be used for ordinary linear regression, factor analysis, and structural equation modelling, it underestimates standard errors because it fails to account for the uncertainty inherent in the missing data which arises from the fact that we have to *estimate* the unobserved ‘true’ values. Moreover, expectation-maximisation is not practical for estimating logistic regression coefficients, as no suitable commercial software is currently available (see Millsap & Maydeu-Olivares 2009 for further details).

A final word of caution on the use of expectation-maximisation is in order. It is important to recognise that the expectation-maximisation algorithm was designed for the purpose of estimating population parameters, *not* for imputing plausible values that can be ‘plugged in’ to replace missing data points and provide researchers with a complete dataset on which subsequent analyses can be carried out directly. When used in this manner, expectation-maximisation effectively results in regression imputation. Researchers should refrain from using expectation-maximisation to create complete datasets as a basis for further data analysis because doing so would yield biased research results (Enders 2010).

Direct maximum likelihood

Direct maximum likelihood is an alternative to expectation-maximisation. Remember that with expectation-maximisation we first ‘fill in’ the missing values and then find the maximum likelihood estimates for the complete-data problem. With direct maximum likelihood, we maximise the likelihood function directly based on parameters from a specified distribution. Although applied social science researchers routinely specify a multivariate normal distribution, other distributions are also possible.

A notable advantage of direct maximum likelihood is its ability to produce unbiased parameter estimates and standard errors under the multivariate normal model (Allison 2002). Parameter estimates are robust to deviations from normality, although standard errors are underestimated. However, both direct maximum likelihood and expectation-maximisation procedures are sensitive to misspecifications of the imputation model. Direct maximum likelihood can be used in linear models (including structural equation modelling, hierarchical linear modelling), yet specialised software or the adaptation of standard software is required to carry out direct maximum likelihood estimation for non-linear analyses (Millsap & Maydeu-Olivares 2009). While detailing the mechanics of direct maximum likelihood is beyond the scope of this paper, we refer the interested reader to Enders (2010) for an excellent and accessible discussion.

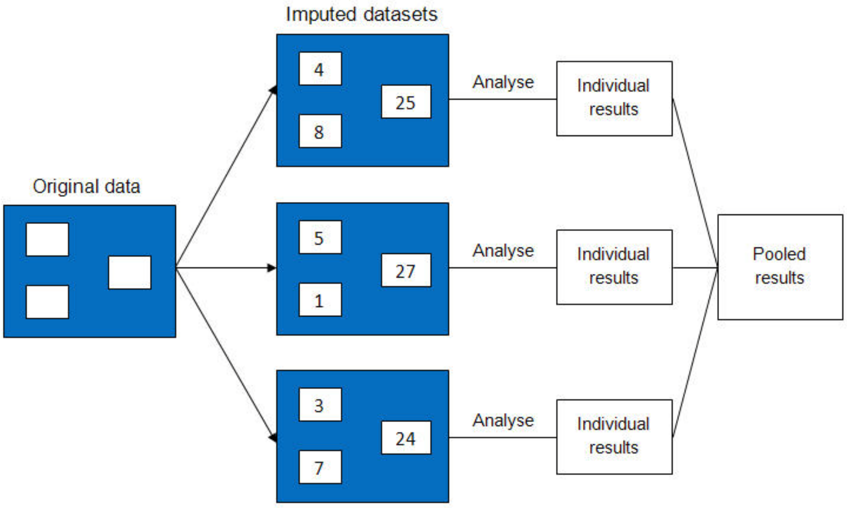
Multiple imputation

Multiple imputation is a modern, general-purpose missing data method that can be applied to any type of data and used with any kind of statistical analysis. The basic process of multiple imputation is straightforward and consists of the following broad steps:

- First, we specify an imputation model to replace missing values with plausible data points. These plausible data points are drawn at random from an assumed underlying distribution of the missing data that is based on the distribution of the observed data. (Usually we assume a multivariate normal distribution, although other distributions are possible.) This process turns the original incomplete dataset into a complete one.
- Adding random variation to the imputation model, we then repeat this process several times to obtain a specified number of complete datasets. Due to the added random variation, each dataset will replace the missing values with a slightly different set of plausible data points.
- Next, we analyse each of the now complete datasets using any statistical analysis method we deem appropriate to answer our substantive research question. This process provides us with a set of results (parameter estimates and standard errors) for each imputed dataset. Given that the imputed values for each dataset are slightly different, the parameter estimates and standard errors should also differ.
- Finally, we use simple arithmetic procedures that are built into standard statistical software packages to pool the different parameter estimates and standard errors into a single set of results. This procedure ensures that the uncertainty inherent in the missing data that arises from the fact that we have to *estimate* the unobserved true values is accounted for by upward-adjusting the pooled standard error estimate.

The generation of several different values for each missing data point is a key characteristic that distinguishes multiple imputation from all other missing data methods. A simple diagram of multiple imputation is given in figure 2. Here, the original data has three missing data points, represented by blank squares. The multiple imputation process creates three imputed datasets in which the missing data points are ‘filled in’ with slightly different plausible values each time. This is done to reflect the uncertainty inherent in the missing data, which arises from the fact that we have to *estimate* the unobserved ‘true’ values. Each imputed dataset is analysed separately before pooling the individual parameter estimates and standard errors into a single set of results. (This is done automatically by statistical software packages that incorporate multiple imputation functionality.)

Figure 2 Simple diagram of the multiple imputation process



To further clarify the mechanics of generating multiple imputed datasets, the following provides a brief conceptual overview of how the process is implemented in the statistical software package SAS. We provide more details and an example of the imputation process in appendix B: A worked example of multiple imputation.

How multiple imputation works in SAS

The multiple imputation framework offers several methods to create plausible data points for missing values. Which method to choose depends on the pattern of the missing data. The three multiple imputation methods available in SAS include:

1 regression method for monotone missing data

2 propensity score method for monotone missing data

3 data augmentation method for arbitrary missing data
(also known as the *Markov Chain Monte Carlo* method).

Given that our study addresses arbitrary missingness, we focus on the data augmentation algorithm (Schafer 1997), which is the default option in SAS.[2] The data augmentation method consists of an iterative two-step algorithm. The first step is called the *imputation step*, and the second the *posterior*[3] *step*.

The imputation step

The imputation step uses stochastic regression imputation (regression imputation with a random error term) to create plausible data points for replacing the missing values. This means that missing values are predicted using regression equations, but each predicted value is then *augmented* with a normally distributed random error term. The reason for adding the error term is to restore lost variability to

[2] The software package R uses a different method for multiple imputation called Multiple Imputation by Chained Equations. Readers are referred to Azur, Stuart and Frangakis (2011) for a description of this alternative method.

[3] The term *posterior* is a reference to the posterior distribution used in Bayesian statistics. A posterior distribution is specified to describe the relative probability of different parameter values. It is the distribution from which the plausible data points are drawn. An excellent description of Bayesian statistics in the context of multiple imputation is available by Enders (2010).

the data and avoid the biases inherent in standard regression imputation (that is, overestimated correlations and R^2 values; underestimated standard errors). Conceptually, we can say that the imputation step simulates a random draw from a set of plausible data points to replace the missing values, conditional on the observed values.

The posterior step

The posterior step takes the 'filled in' values from the previous imputation step and calculates the mean vector and covariance matrix of the now complete dataset. It then randomly perturbs the values for the mean vector and covariance matrix and passes them on to the next imputation step. This next imputation step uses these perturbed values to generate a *new* set of regression equations, which in turn leads to different 'filled in' plausible data points and, therefore, to a new complete dataset in which the replacement values are slightly different. This process is repeated multiple times, depending on the number of requested imputations. (For additional details on the posterior step, along with a step-by-step example, see appendix B: A worked example of multiple imputation.)

Pooling the results

We have already mentioned that once the desired number of imputed datasets has been created, we analyse each dataset individually with whichever statistical method we prefer to answer our substantive research question. In a final step, we now pool (that is, combine) parameter estimates and standard errors obtained from analysing each respective imputed dataset into a single set of results. Parameter estimates are pooled in one simple step by averaging the results from all separately analysed datasets. The pooling of standard errors consists of (1) calculating the *within*-imputation variance by averaging standard errors over all imputed datasets, (2) calculating the *between*-imputation variance of the parameter estimates over all imputed datasets, and (3) taking the square root of the total variance of the parameter estimate. For completeness, we summarise the pooling process in arithmetic terms in table 3. However, note that statistical software packages carry out this process automatically for all standard analysis methods.

Table 3 Pooling procedure for multiple imputation parameter estimates and standard errors

Step	Formula	
1. Pooled parameter estimate	$\bar{Q} = \frac{1}{m}\sum_{i=1}^{m} \hat{Q}_i$	where m is the number of imputations and $\hat{Q}_i$ is the parameter estimate from the *i*-th imputed dataset
2. Pooled standard error		
a. Within-imputation variance	$\bar{U} = \frac{1}{m}\sum_{i=1}^{m} \hat{U}_i$	where $\hat{U}_i$ is the variance estimate from the *i*-th imputed dataset, and *m* is the number of imputations
b. Between-imputation variance	$B = \frac{1}{m}\sum_{i-1}^{m} (\hat{Q}_i - \bar{Q})^2$	
c. Total imputation variance	$T = \bar{U} + \left(1 + \frac{1}{m}\right) B$	
d. MI standard error	S.E. = $\sqrt{T}$	

Advantages of multiple imputation

Multiple imputation has become tremendously popular among researchers, particularly in the fields of medicine and bio-statistics. This popularity is based on a number of key advantages multiple imputation can offer over other missing data methods.

Suitable when data are missing at random

One important shortcoming of listwise deletion, mean substitution, and other basic approaches is that these methods produce unbiased parameter estimates only when data are *missing completely at random*. Similar to maximum likelihood-based methods, multiple imputation yields unbiased parameter estimates under the much more realistic *missing at random* mechanism.

In addition to producing unbiased parameter estimates when data are *missing at random*, multiple imputation yields correct standard errors that incorporate the uncertainty inherent in the missing data that arises from the fact that we have to *estimate* the unobserved true values. This added uncertainty is reflected in larger standard errors for each parameter estimate. Case deletion and single imputation methods fail to account for this additional uncertainty around parameter estimates. Given constant sample size, methods that ignore the uncertainty in the missing data will underestimate standard errors and *p*-values and increase the Type I error risk (Schafer & Olsen 1998).

Appropriate for mixed-variable datasets

Large-scale datasets in the social sciences usually contain numerous binary and categorical variables. However, single imputation methods have been developed for continuous multivariate-normal data and are inefficient when used to address binary and categorical missing values. While existing maximum-likelihood methods have been adapted for use with categorical missing data (see Lipsitz & Ibrahim 1996), these adaptations require a high level of technical expertise and are difficult to implement in practice. A major advantage of multiple imputation in the context of social science research with large-scale datasets is that researchers can use a single, straightforward process to impute variables with continuous as well as non-continuous data.

Multiple imputation offers maximum flexibility

Maximum likelihood methods integrate the imputation and analysis of data into one overall process. Since these methods work primarily with linear models, researchers are limited to certain statistical models when trying to answer their substantive research question. Multiple imputation completely separates the imputation from the data analysis process, thereby giving researchers complete flexibility over the post-imputation statistical models they wish to use.

How well do different methods perform?

This section has reviewed a number of missing data methods that are commonly used in the social sciences. So how well do different methods cope with missing data in real-world social science datasets? We answer this question in the following section using data from a longitudinal survey and an administrative collection.

Performance test

Overview

The relative performance of different missing data methods can be evaluated in terms of their ability to yield analysis results (that is, parameter estimates and their associated standard errors) from an incomplete dataset that closely resemble the results that would have been obtained had the dataset been complete. Based on this idea, we followed a five-step process to ascertain how well select missing data methods perform under different mechanisms and varying levels of missingness. We began by choosing a complete sample of data from two social science collections. In a second step, we deleted values from each of the two samples in a controlled manner so as to create a series of new samples, each featuring a predetermined level of missingness and operating under a specified missing data mechanism. We subsequently used a standard statistical model to analyse the two original complete samples in order to obtain a set of 'true' analysis results. Using different missing data methods, the same model was then run on each of the newly created incomplete samples. Finally, we compared the results obtained for each incomplete sample with those of the respective complete dataset. The following sections illustrate each of the five steps in more detail.

Step 1: Selecting the complete samples

We used data from the 2003 cohort of the Longitudinal Surveys of Australian Youth (LSAY) and the 2009 National Vocational Education and Training Provider Collection (VET Collection) for our performance assessment. Including samples from two different social science datasets in our examination allowed us to cross-check performance test results across datasets with different variable compositions. Specifically, the LSAY sample contained more continuous variables, whereas the VET Collection sample contained more binary variables.

LSAY

LSAY is a nationally representative survey that tracks young people from the ages of 15 to 25 as they move from school into further study and work. We randomly selected a sample of 5000 cases with complete data on four predictors and one outcome variable to predict completion of the senior secondary certificate of education from a respondent's sex, occupational aspirations, socioeconomic status, and mathematics achievement. We standardised all continuous variables to facilitate the comparability of results. Table 4 provides descriptive data for the complete LSAY sample.

Table 4 Descriptive data for the complete LSAY sample

Variable	Coding	M	SD
Sex	Male = 0; female = 1	.51	.500
Occupational aspirations[a]	Continuous	.00	1.001
Socioeconomic status[b]	Continuous	.00	1.002
Mathematics achievement[c]	Continuous	.00	1.000
SSCE completion status[d]	Yes = 0; No = 1	.15	.354

Notes: a Occupational aspirations were measured using the International Socio-Economic Index of Occupational Status (ISEI, Ganzeboom et al. 1992). Higher scores indicate higher levels of expected occupational status.

b Socioeconomic status was measured using the Index of Economic, Social, and Cultural Status (ESCS), a composite measure of parental occupation, parental education, and home possessions.

c Mathematics achievement was measured using the first of five plausible values from the 2003 Program for International Student Achievement (PISA), which forms the base year of the 2003 LSAY cohort.

d The proportion of students who do not complete the senior secondary certificate of education is underestimated in LSAY because respondents who do not complete secondary education are more likely to drop out of the survey. LSAY weights only partially account for this attrition bias.

VET Collection

The VET Collection is a large administrative dataset that provides information on vocational training programs from government-funded and privately operated training providers. Its main purpose is to measure vocational course completion rates and other indicators of the vocational education and training system. We sampled all 3542 cases from a total of 4943 vocational bridging course[4] participants who had complete data on five predictors and one outcome variable to predict completion of the bridging course from a respondent's age, disability status, completion status of the senior secondary certificate of education, completion status of a vocational certificate III,[5] and non-English speaking background. Table 5 provides frequency distributions for the complete VET Collection sample.

Table 5 Frequency distributions for the complete VET Collection sample

Variable	Values	n	%
Age	16–67	N/A	N/A
Disabled	0 = No 1 = Yes	3471 71	98.0 2.0
Completed senior secondary certificate of education	0 = No 1 = Yes	2841 728	79.4 20.6
Completed cert. III	0 = No 1 = Yes	2056 1486	58.0 42.0
Non-English speaking background	0 = No 1 = Yes	1882 1660	53.1 46.9
Dropped out of bridging course	0 = No 1 = Yes	3068 474	86.6 13.4

[4] Bridging courses include a broad range of introductory vocational courses, such as office administration, bookkeeping etc.

[5] A certificate III is a vocational education and training sector accreditation within the Australian Qualifications Framework. It allows individuals to perform a range of skilled operations in the trades or other occupations.

Step 2: Creating samples with different missing data mechanisms and levels of missingness

We created a series of incomplete samples under different missing data mechanisms and levels of missingness by deleting values from the complete LSAY and VET Collection samples in a controlled manner. Specifically, we used straightforward probabilistic deletion to create incomplete samples under *missing completely at random* (MCAR), and a series of logistic regression models to create incomplete samples under *missing at random* (MAR).[6] It is important to note that the sole purpose of these logistic models was to determine which values to delete from the complete samples. As such, they were *entirely unrelated* to the logistic regression models we used later on to analyse the samples.

Imposing an MCAR mechanism

We recall that under missing completely at random the probability of a given value being missing on variable *X* does not depend on the values of any other variable in the dataset. Suppose we want to model the probability that the predictor *mathematics achievement* will be missing for some respondents in the LSAY sample. Since values for that predictor need to be deleted completely at random, we want the probability of deletion to be constant (that is, the same for every respondent).

Suppose the probability that a respondent's mathematics achievement score is missing is fixed at 7%. To apply the calculated probability to the data, we create a random variable, *u*, from a uniform [0,1] distribution (that is, a random number between 0 and 1). We then delete the mathematics achievement score if *u* is less than .07, which is equivalent to the score having a 7% probability of deletion. An illustration of this deletion process is provided in table 6.

Table 6 Illustration of data deletion process for MCAR

Respondent	$Pr_{(\text{MATHS MISSING})}$	u ~ [0,1]	Delete MATHS if $u \leq Pr$
1	.07	.315	Keep
2	.07	.724	Keep
3	.07	.012	Delete

Respondent 3 above would have their mathematics achievement score deleted because *u* is less than the given probability, whereas respondents 1 and 2 would retain their scores. On average, seven in every 100 respondents would have their scores deleted. Note that each respondent has an *equal* chance of having their mathematics achievement score deleted, meaning that the probability of deletion does not depend on any other variable in the dataset.

Here, we chose an arbitrary value for the probability of mathematics achievement being missing. This value was then adjusted to give the desired overall percentage of missingness. Full details of the probabilities used for all of our *missing completely at random* mechanisms can be found in tables C5 and C9 of appendix C.

[6] A *missing not at random* mechanism was not included in our analysis because the complexities associated with non-random missingness were beyond the scope of this paper. The interested reader is referred to a recent simulation study by Marshall et al. (2010) which found that under *missing not at random* all tested methods, including multiple imputation, performed poorly with 25% or more overall missingness.

Imposing a MAR mechanism

Under *missing at random* the probability of a value being missing on variable *X* depends on the values of one or more observed covariates, but must *not* depend on the value of *X* itself. To illustrate, in a dataset with three predictors X_1, X_2, X_3, and one outcome variable *Y*, the probability that X_3 is missing for a given respondent can be related to that respondent's values of X_1, X_2, and *Y*, but must not depend on the value of X_3. The probability of deletion is no longer constant for every observation. Therefore, we need to use a logit model to impose the *missing at random* mechanism. The logit model consists of a constant term *plus* additional terms.

To illustrate, suppose we would like the probability of X_3 (a respondent's mathematics achievement score) being missing to depend on:

- *Y* (WHETHER OR NOT THE PERSON COMPLETED SCHOOL)
- *X1* (THE PERSON'S OCCUPATIONAL ASPIRATION SCORE)
- *X2* (THE PERSON'S SOCIOECONOMIC STATUS)

and for this probability to *increase* if X_1 is already missing. These rules can be modelled with the following logit function,

$$\text{logit}[\Pr_{(\text{MATHS MISSING})}] = \alpha_0 + \alpha_1 Y + \alpha_2 X_1 + \alpha_3 X_2 + \alpha_4 MX_1,$$

where α_0, α_1, α_2, α_3, and α_4 are constants, and MX_1 is a missingness indicator for predictor X_1 which equals 1 if X_1 is missing, and 0 if it is observed. The coefficients α_0 through α_4 were adjusted until we obtained the desired level of missingness. A parallel goal was to choose 'sensible' values for the logit coefficients to implement a realistic *missing at random* mechanism. A detailed example of how 'sensible' values for the logit coefficients are chosen is provided in *appendix C*. Once the coefficients are set, the probabilities of deletion are found using the inverse logit function, and the remainder of the deletion process for *missing at random* is then identical to that of *missing completely at random* (see appendix C for details).

Note on our implementation of MAR

Strictly speaking, a *missing at random* mechanism is not allowed to depend on any unobserved variables. Since the subset of observed variables could be different for every respondent (for example, one respondent might only have values for sex, occupational aspirations, and mathematics achievement, while another might only have values for occupational aspirations and socioeconomic status), we would need to define a separate *missing at random* mechanism for each scenario.

Many simulation studies circumvent this problem by having the *missing at random* mechanism depend only on variables that are fully observed for all respondents. Such an approach would not be appropriate for our purposes, as we only have one fully observed variable, *Y*, and a number of predictors on which we want to impose missingness. Although we could let the probability of each predictor being missing depend *only* on *Y*, this approach would be overly simplistic and unrealistic.

We tackled this issue by implementing a 'hybrid' *missing at random* mechanism, whereby the missingness for a given variable *X* could *not* depend on that predictor itself, but *could* depend on all other variables, some of which may be unobserved for a particular respondent. While our implementation of *missing at random* somewhat deviates from the 'text book' definition, we argue that it better reflects the reality of missing value dependencies. As Graham (2007) argues, data are rarely strictly *missing at random* or strictly *missing not at random*, but somewhere on a continuum between the two.

Predictors with imposed missingness

Following the above process for imposing different missing data mechanisms, we deleted four predictors in LSAY, including sex, occupational aspirations, socioeconomic status, and mathematics achievement. Likewise, we deleted three predictors in the VET Collection, including completion status of the senior secondary certificate of education, completion status of a vocational certificate III, and non-English speaking background.

Setting levels of missingness

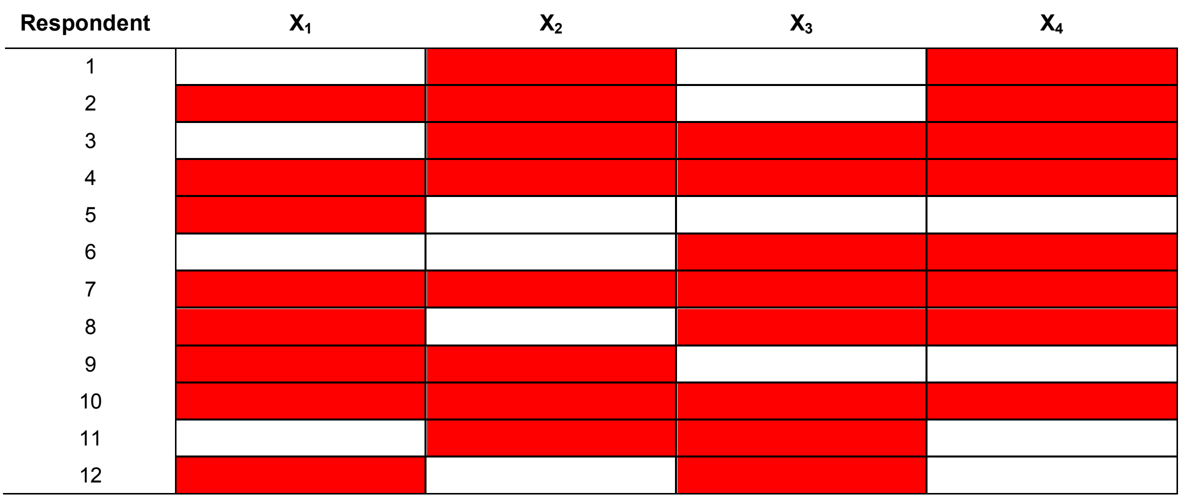
For each missing data mechanism, we created three different levels of missingness, including a moderate level of 25%, a high level of 50%, and a 'realistic' level that reflected the actual missingness in the original collections. This actual level was 17% across the four incomplete predictors in LSAY and 30% across the three incomplete predictors in the VET Collection. The resulting combinations of mechanisms and levels of missingness are outlined in table 7.

Table 7 Combinations of missing data mechanisms and levels of missingness

Mechanism	Level of missingness (%)	
	LSAY	VET Collection
Missing completely at random	17	25
	25	30
	50	50
Missing at random	17	25
	25	30
	50	50

In table 7, a level of missingness of 25% means that '25% of respondents have at least one data value missing'. Note that it does *not* mean that 25% of values are deleted from each predictor. Since values are deleted from multiple predictors, predictors can have overlapping missingness (see figure 3), meaning that we only need to delete a small percentage of values from each column to arrive at 25% missing overall.

Figure 3 Illustration of overlapping missing values across predictors



Respondent	X_1	X_2	X_3	X_4
1		🟥		🟥
2	🟥	🟥		🟥
3		🟥	🟥	🟥
4	🟥	🟥	🟥	🟥
5	🟥			
6			🟥	🟥
7	🟥	🟥	🟥	🟥
8	🟥		🟥	🟥
9	🟥	🟥		
10	🟥	🟥	🟥	🟥
11		🟥	🟥	
12	🟥		🟥	

Note: Missing values are represented by a blank rectangle; complete values are represented by a coloured rectangle.

Let us consider figure 3, which illustrates a dataset with 12 records and four variables. Even though only 25% of values are missing from each column, 75% of records (9 out of 12) have missing data overall. Thus, to achieve a specified level of missingness (for example, 25%) we would delete values across all predictors such that 25% of the records have *at least one* predictor missing by adjusting the α_i values for each predictor's logit function.

Following this example, we generated 1000 slightly different incomplete samples for each combination of missing data mechanism and level of missingness listed in table 7. This was done to ensure that our conclusions about the performance of missing data methods would be robust (that is, not vary greatly from one possible incomplete sample to the next). For each set of 1000 runs, variability was introduced by applying different seed values to generate the uniform random variable *u* that was used to control the deletion process.

Verifying missing data mechanisms

We conducted Little's MCAR test to check our newly created incomplete samples. Results from the test for the LSAY and VET Collection samples are summarised in table 8.

Table 8 Results from Little's MCAR test for LSAY and VET Collection samples

Mechanism		**LSAY**		**VET Collection**	
	% Missing	χ^2	p	χ^2	p
Missing completely at random	17	67.7	.180	N/A	N/A
	25	59.7	.558	27.5	.595
	30	N/A	N/A	29.8	.790
	50	59.2	.879	29.1	.514
Missing at random	17	605.8	<.001	N/A	N/A
	25	1165.8	<.001	1181.4	<.001
	30	N/A	N/A	1334.4	<.001
	50	1995.6	<.001	1310.2	<.001

We recall that Little's MCAR test rejects the null hypothesis of completely random missingness if the missing values do not represent a true random sample of the observed data. Results from table 8 confirm that the null hypothesis of complete random missingness is *not* rejected for our MCAR samples, but *is* rejected for our MAR samples. This confirms that the MCAR datasets we created had indeed *completely* random missingness, as opposed to our MAR datasets.

Step 3: Analysing the complete samples

We used logistic regression analysis as the standard statistical model for comparing the performance of select missing data methods. Our decision to choose logistic regression over alternative procedures (for example, linear regression, ANOVA etc.) was based on the fact that performance comparisons of missing data methods using linear models are readily available in the literature (see Olinsky, Chen & Harlow 2003; Schafer & Graham 2002). Given that studies in the social sciences often focus on binary outcomes (for example, completion of the senior secondary certificate of education, enrolment in tertiary education, completion of apprenticeships or vocational training modules, full-time employment status, receipt of welfare benefits), we selected binary logistic regression analysis for our performance test.

We recall that the overall objective of analysing the complete samples is to create a set of 'true' baseline results against which to compare results from the different incomplete samples. The specific baseline results we were interested in included regression coefficients and their associated standard errors. To generate these baseline results for the complete LSAY sample, we predicted the probability of completing the senior secondary certificate of education, given an individual's sex, occupational aspirations, socioeconomic status, and mathematics achievement. We used the following logit model[7],

$$\text{logit } [\Pr_{(\text{COMPLETING SSCE})}] = \alpha_0 + \alpha_{1(\text{SEX})} + \alpha_{2(\text{OCC\_ASP})} + \alpha_{3(\text{SES})} + \alpha_{4(\text{MATH})}$$

where α_0 is the intercept and α_1 through α_4 are the regression coefficients for each of the four predictors. Baseline results for the complete LSAY sample are provided in table 9. Note that all predictors are significant.

Table 9 Logistic regression results for the complete LSAY sample

	β	SE	Wald χ^2	df	p
Intercept	1.981	.069	834.09	1	<.001
Sex	.520	.087	36.04	1	<.001
Occupational aspirations	.598	.044	185.46	1	<.001
Socioeconomic status	.179	.045	15.98	1	<.001
Mathematics achievement	.670	.047	204.71	1	<.001

For the VET Collection sample, we used a similar logit model to predict the probability of dropping out of a vocational bridging course, given an individual's disability status, age, completion of the senior secondary certificate of education, completion of a vocational certificate III, and non-English speaking background status, such that

$$\text{logit } [\Pr_{(\text{DROPOUT})}] = \beta_0 + \beta_{1\ (\text{DISABLED})} + \beta_{2\ (\text{AGE})} + \beta_{3\ (\text{SSCE})} + \beta_{4\ (\text{CERT III})} + \beta_{5\ (\text{NESB})}$$

where β_0 is the intercept and β_1 through β_5 are the regression coefficients for each of the five predictors. Baseline results for the complete VET Collection sample are provided in table 10. Note that all predictors are significant.

Table 10 Logistic regression results for the complete VET Collection sample

	β	SE	Wald χ^2	df	p
Intercept	-.729	.187	15.16	1	<.001
Age	1.500	.271	30.74	1	<.001
Disabled	-.028	.007	13.93	1	<.001
Completed SSCE	.618	.117	27.87	1	<.001
Completed cert. III	-.494	.114	18.72	1	<.001
Non-English speaking background	-1.360	.124	120.30	1	<.001

[7] Note that the logistic regression models used to analyse the LSAY and VET Collection samples are for illustration purposes only. As such, they are intentionally kept very basic and should not be used to draw substantive conclusions.

Step 4: Analysing the incomplete samples with different missing data methods

Using several different missing data methods, we fit the logistic regression models outlined in the previous step to each of the incomplete samples. The missing data methods we used include *listwise deletion*, *constant replacement*, and *multiple imputation*. We chose listwise deletion because it is among the most widely practised approaches to handling missing data in the social sciences. Listwise deletion is the default option in standard statistical software packages. Constant replacement was included because substituting missing data with the variable mean or mode is a simple option to address the issue. For the LSAY sample, constant replacement was implemented by using mean substitution for continuous predictors and mode substitution for binary predictors. For the VET Collection sample, only mode substitution was used, given that all predictors with missing values were binary. Finally, we included multiple imputation in our study because it is considered one of two 'gold standard' methods for addressing missing data. The inclusion of multiple imputation as a standard option in many general-purpose statistical software packages makes the method readily accessible for social scientists with various levels of statistical expertise. We implemented multiple imputation with ten and 100 imputed datasets to gauge potential performance gains from increasing the number of imputations. Multiple imputation was carried out in the software package SAS. (The code used to carry out multiple imputation in the LSAY and VET Collection samples is provided in appendix F: SAS code.) The second 'gold standard' method, direct maximum likelihood, was excluded because we only tested methods that can be easily implemented using standard statistical software.[8]

Step 5: Comparing results

As a final step, we compared the logistic regression results for each of the complete samples with those from the incomplete samples for every combination of method, mechanism, and level of missingness. Specifically, we assessed regression coefficients and their associated standard errors for every predictor in the logistic regression model. Performance was assessed in terms of the deviation from the 'true' complete-sample results.

[8] Direct maximum likelihood requires the use of specialised software or adaptations of standard software for use with logistic regression analysis. As such, the use of direct maximum likelihood may present a challenge for less experienced social science researchers. Messer and Natarajan (2008) carried out a simulation study using direct maximum-likelihood-based imputation for logistic regression. The authors found that direct maximum-likelihood and multiple imputation performed equally well under realistic missing data scenarios.

Results

The following results illustrate the extent to which regression coefficients and associated standard errors deviated from those of the complete samples under each of the missing data methods.

Regression coefficients for LSAY

We computed the percentage deviation in regression coefficients from the complete sample by predictor for each combination of missing data mechanism and level of missingness. These estimates are stable because they represent mean results over 1000 simulation runs (that is, over 1000 randomly drawn samples, as was explained in the *Setting levels of missingness* section). Table 11 provides a summary of our results for regression coefficients. Deviations in regression coefficients in excess of 25% are highlighted in bold.

Table 11 Percentage deviation in regression coefficients from the complete LSAY sample

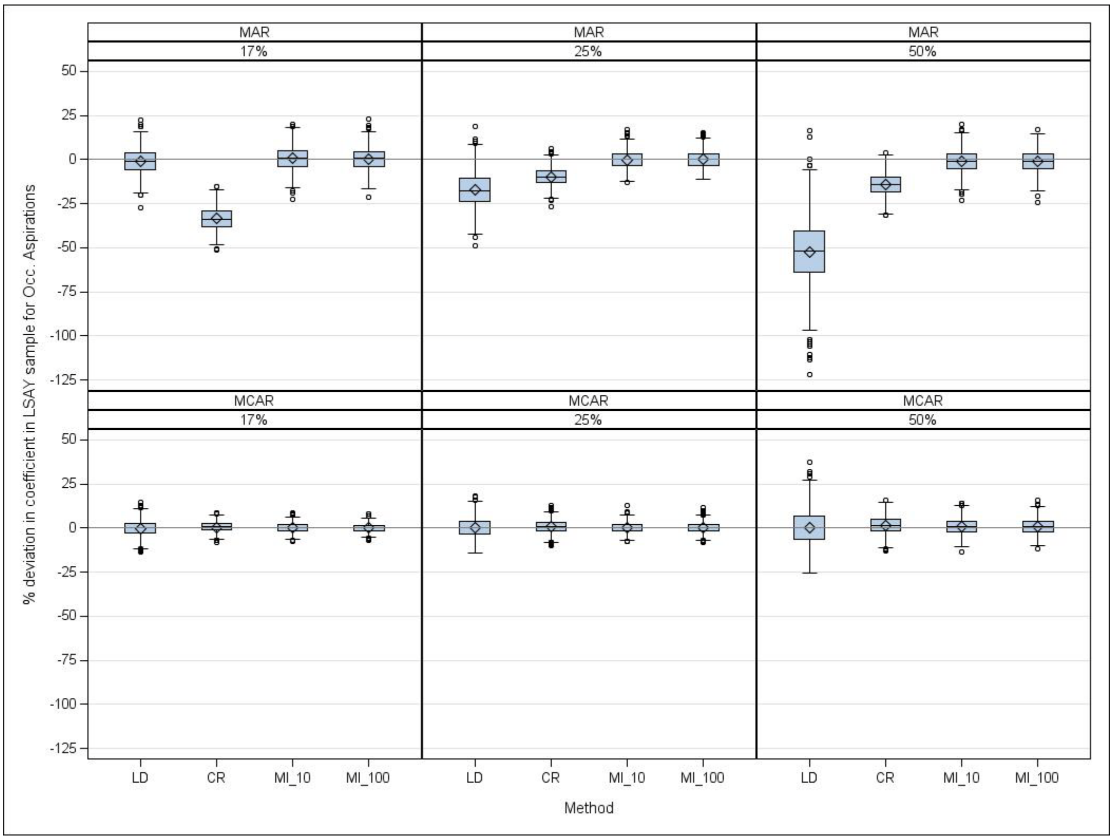
Predictor	Method	MCAR			MAR		
		17	25	50	17	25	50
Sex	Listwise deletion	0.03	0.36	0.08	-0.18	-3.08	-17.96
	Constant replacement	-5.27	-7.44	-16.64	12.58	***-84.57***	***-164.90***
	Multiple imputation (10)	0.68	1.31	2.22	1.57	-2.20	-4.24
	Multiple imputation (100)	0.79	1.30	2.23	1.62	-2.21	-4.21
Occ. asp.	Listwise deletion	-0.10	0.27	0.54	-0.82	-17.05	***-52.38***
	Constant replacement	0.54	0.82	1.59	***-33.59***	-9.74	-14.07
	Multiple imputation (10)	0.22	0.38	1.01	0.60	-0.10	-0.94
	Multiple imputation (100)	0.19	0.41	1.04	0.49	-0.02	-1.00
SES	Listwise deletion	-0.27	-0.56	-0.68	***-30.15***	***-74.13***	***-87.73***
	Constant replacement	7.09	10.09	22.63	16.30	***-33.61***	***-30.18***
	Multiple imputation (10)	1.85	2.28	6.59	-1.76	-15.75	-4.99
	Multiple imputation (100)	1.78	2.42	6.54	-1.74	-15.09	-4.44
Maths	Listwise deletion	0.01	-0.01	0.03	-6.65	-4.51	4.44
	Constant replacement	-1.58	-2.03	-3.94	11.86	***-30.89***	-20.42
	Multiple imputation (10)	-0.33	-0.37	-1.07	-1.30	-11.23	-5.47
	Multiple imputation (100)	-0.32	-0.41	-1.08	-1.29	-11.29	-5.42

For samples where missingess was completely at random, we found the regression coefficients resulting from listwise deletion and multiple imputation to deviate only marginally from those produced by the complete sample. This outcome was unsurprising, since both methods are known to be efficient under MCAR. However, we also wish to remind the reader that listwise deletion leads to a loss of statistical power regardless of its good performance in relation to coefficient bias under MCAR.

Our much more realistic MAR samples clearly demonstrate the relative performance advantage of multiple imputation over the two basic methods. Listwise deletion and constant replacement led to very large deviations across most predictors. In comparison, multiple imputation performed consistently well under MAR, even when 50% of cases had incomplete data. We note that there was no difference between creating ten versus 100 imputed datasets. We ascribe this to the large sample size we used for simulation. For small samples, a larger number of imputed datasets would likely increase estimation accuracy (see Graham, Olchowski & Gilreath 2007).

A graphical representation of results for individual variables provides further important insights. For example, figure 4 presents regression coefficient results for *occupational aspirations*. Graphs for the coefficients of all other predictors show a similar pattern and are provided in appendix D: Regression coefficients for LSAY and the VET Collection.

Figure 4 Percentage deviation in regression coefficients from the complete LSAY sample for *occupational aspirations*



Note: LD = listwise deletion; CR = constant replacement; MI_10 = multiple imputation with 10 imputed datasets; MI_100 = multiple imputation with 100 imputed datasets.

In addition to reflecting the deviation results outlined in table 10, figure 4 illustrates an additional dimension, namely the variability of estimation results over 1000 simulation runs. Clearly, the loss of statistical power from listwise deletion causes the largest spread in estimates for the predictor *occupational aspirations*. This means that when using a single sample (as opposed to the 1000 possible random samples that were drawn for this simulation study), the regression coefficient for *occupational aspirations* could be either highly overestimated or highly underestimated.

Regression coefficients for the VET Collection

In the VET Collection sample, all but one predictor were binary, and missingness was imposed on binary predictors only. Given the strong departure from the multivariate normal model the performance of multiple imputation was of particular interest to us. Table 12 provides the overall percentage deviation in regression coefficients over 1000 logistic regression runs for each combination of missing data mechanism and level of missingness. Deviations in regression coefficients in excess of 25% are highlighted in bold.

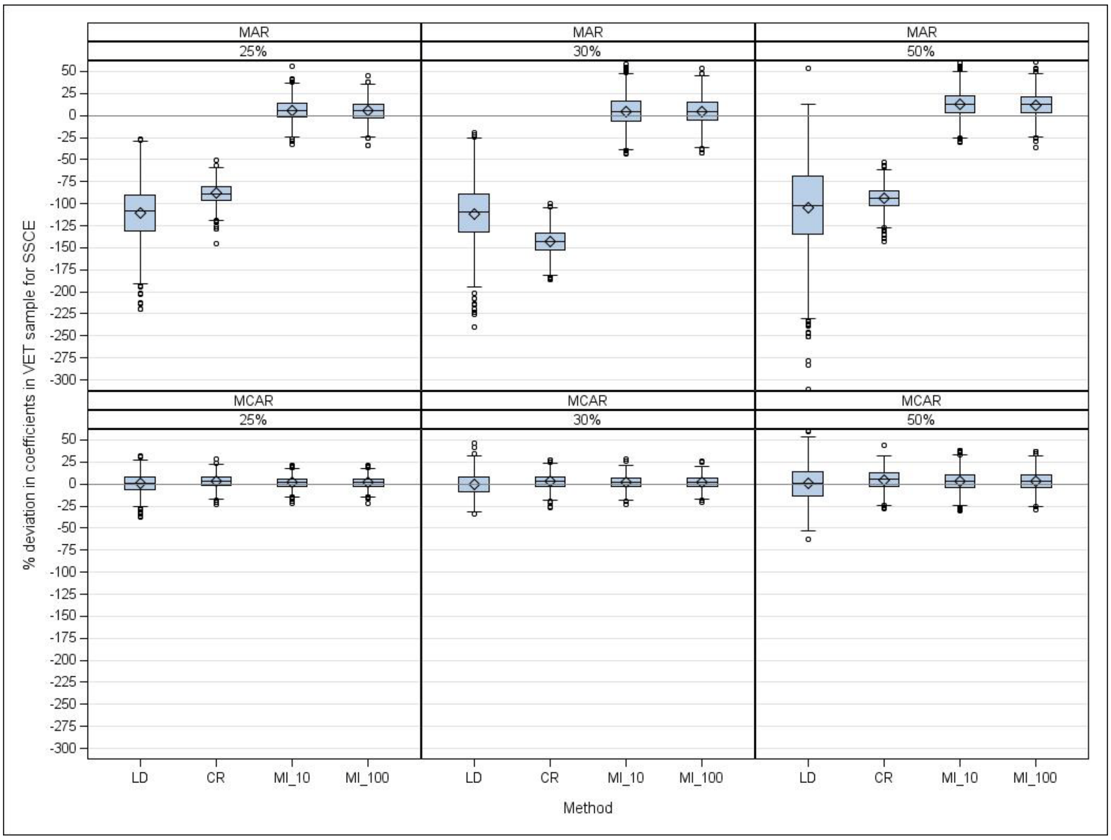
Table 12 Percentage deviation in regression coefficients from the complete VET Collection sample

Predictor	Method	MCAR			MAR		
		25	30	50	25	30	50
Age	Listwise deletion	-0.55	-0.45	-0.66	***-37.75***	***-63.17***	***-72.55***
	Constant replacement	6.27	7.99	14.48	***48.70***	***59.95***	***48.07***
	Multiple imputation (10)	-0.26	-0.25	-0.25	-7.51	-22.90	-5.90
	Multiple imputation (100)	-0.30	-0.28	-0.31	-7.39	-23.08	-5.46
Disabled	Listwise deletion	0.02	0.02	0.16	3.42	5.51	0.93
	Constant replacement	1.64	2.20	3.80	8.30	12.70	10.22
	Multiple imputation (10)	0.22	0.43	0.59	4.68	5.03	7.53
	Multiple imputation (100)	0.19	0.42	0.55	4.60	5.06	7.45
SSCE	Listwise deletion	0.62	-0.37	0.46	***-110.81***	***-111.49***	***-104.23***
	Constant replacement	3.04	2.94	4.88	***-88.36***	***-143.19***	***-94.15***
	Multiple imputation (10)	1.62	1.70	3.41	5.78	4.69	12.63
	Multiple imputation (100)	1.61	1.62	3.49	5.28	4.44	12.09
Cert. III	Listwise deletion	-0.55	-0.17	-0.49	***105.35***	***122.65***	***149.19***
	Constant replacement	3.47	4.34	7.88	***-100.46***	***-115.82***	***-93.18***
	Multiple imputation (10)	-0.07	0.38	0.54	23.72	***32.29***	14.01
	Multiple imputation (100)	0.01	0.60	0.74	23.55	***32.52***	13.60
NESB	Listwise deletion	-0.08	-0.18	-0.41	-21.77	***-77.50***	-22.99
	Constant replacement	4.66	5.85	9.98	-18.76	***-80.51***	***-31.63***
	Multiple imputation (10)	0.91	1.26	2.08	3.56	-19.17	5.92
	Multiple imputation (100)	0.91	1.22	2.13	3.41	-19.51	6.00

Results from the VET Collection samples show a very similar pattern to those from LSAY. Under MCAR, listwise deletion and multiple imputation performed well, whereas constant replacement somewhat overestimated the ‘true’ coefficients. Under MAR, multiple imputation vastly outperformed both basic methods by producing much smaller deviations on all but one predictor (*disabled*). The predictor *non-English speaking background* also shows that multiple imputation can, in some instances, produce quite sizeable deviations from the ‘true’ coefficient (that is, deviations in excess of 30%). Nonetheless, in relative terms the multiple imputation coefficient estimates for that predictor were still much better than those produced by either listwise deletion or constant replacement. Finally, we again saw no benefit from specifying a high versus a low number of imputations.

We use the predictor *senior secondary certificate of education* to illustrate the variability of estimation results over 1000 simulation runs in figure 5. Graphs for the coefficients of all other predictors show a similar pattern and are provided in appendix D: Regression coefficients for LSAY and the VET Collection.

Figure 5 Percentage deviation in regression coefficients from the complete VET Collection sample for *senior secondary certificate of education*



Note: LD = listwise deletion; CR = constant replacement; MI_10 = multiple imputation with 10 imputed datasets; MI_100 = multiple imputation with 100 imputed datasets.

As we can see from figure 5, the loss of statistical power inherent in listwise deletion has a strong detrimental effect on the variability of the estimated coefficients under MAR, even with moderate levels of missingness. The figure further highlights the positive performance of multiple imputation compared with both basic methods. Under MAR, the two basic methods greatly underestimate the impact of completing senior secondary education on dropping out of a vocational bridging course.

Standard errors for LSAY

As with regression coefficients, we computed the per cent deviation in standard errors over 1000 runs for each predictor. Results are summarised in table 13.

Table 13 Percentage deviation in standard errors from the complete LSAY sample

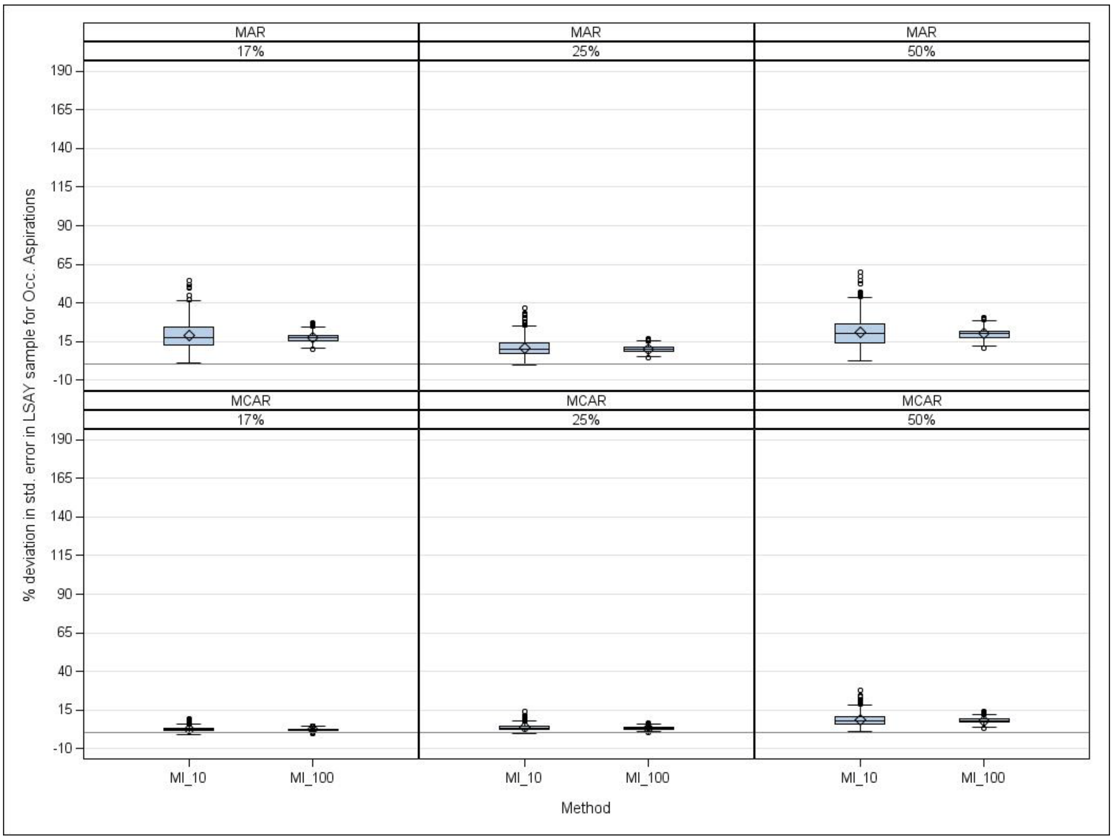
Predictor	Method	MCAR			MAR		
		17	25	50	17	25	50
Sex	Listwise deletion	10.23	15.48	40.30	22.10	42.69	109.06
	Constant replacement	-0.93	-1.26	-1.96	-0.49	-3.31	3.95
	Multiple imputation (10)	2.91	4.28	10.83	1.59	12.64	34.75
	Multiple imputation (100)	2.67	3.97	10.05	1.48	11.84	32.47
Occ. asp.	Listwise deletion	10.12	15.37	40.26	23.15	44.89	129.22
	Constant replacement	0.49	0.82	2.73	0.02	-7.50	-4.25
	Multiple imputation (10)	2.35	3.50	8.48	18.69	10.90	20.66
	Multiple imputation (100)	2.17	3.28	8.00	17.27	9.98	19.82
SES	Listwise deletion	10.25	15.45	40.35	24.28	45.52	126.18
	Constant replacement	1.10	1.68	4.39	-1.68	-5.65	-2.14
	Multiple imputation (10)	2.31	3.46	8.54	1.82	15.55	31.98
	Multiple imputation (100)	2.12	3.22	7.65	1.60	14.47	30.01
Maths	Listwise deletion	10.15	15.42	39.97	21.61	39.65	108.61
	Constant replacement	0.46	0.75	1.74	-1.10	-6.35	-2.01
	Multiple imputation (10)	2.66	3.95	8.43	2.38	13.80	23.26
	Multiple imputation (100)	2.47	3.65	8.00	2.20	13.27	22.31

As a basic missing data method, listwise deletion does not properly reflect the uncertainty inherent in having to estimate the missing data. We would thus expect underestimated standard errors similar to those produced by constant replacement. Instead, we see in table 13 that standard errors for listwise deletion are highly inflated. While this may seem paradoxical at first, we recall that listwise deletion effectively reduces sample size, which results in a loss of statistical power. The standard errors for listwise deletion have thus to be considered relative to their respective level of missingness and the number of records left *after* discarding incomplete cases. It is this reduction in statistical power that leads to drastically overestimated standard errors *relative to those of the complete dataset*.

Similar to listwise deletion, constant replacement does not properly reflect the uncertainty inherent in the missing data. This uncertainty is an expression of our reduced confidence in the parameter estimates, given that we have to *estimate* the missing values. However, in contrast to listwise deletion, constant replacement does not result in standard error inflation relative to the complete sample due to reduced sample size. Instead, we observe in table 13 that constant replacement routinely underestimates standard errors when data are *missing at random*. Underestimated standard errors are problematic because they increase the Type I error risk (that is, rejecting the null hypothesis of no difference/relationship when in fact it is true).

Multiple imputation reflects the uncertainty inherent in the missing data by taking into account the within-imputation and between-imputation variance through the pooling process, as was described in the section, ‘How multiple imputation works in SAS’. The pooling process results in a proper upward adjustment of standard errors. Figure 6 demonstrates this upward adjustment of standard errors for the predictor *occupational aspirations*. The adjustment is relative to the zero line, which represents the ‘true’ standard error for *occupational aspirations* based on the complete sample. Graphs for the standard errors of all other predictors in the LSAY samples show a similar pattern and are provided in appendix E: Multiple imputation standard errors for LSAY and the VET Collection. Note that standard errors for the basic methods have been omitted from the figure, given that they are incorrect relative to those from the complete sample, based on theoretical grounds we have explained above.

Figure 6 Percentage deviation in standard errors from the complete LSAY sample for *occupational aspirations* using multiple imputation



Note: LD = listwise deletion; CR = constant replacement; MI_10 = multiple imputation with 10 imputed datasets; MI_100 = multiple imputation with 100 imputed datasets.

What is particularly interesting in the context of multiple imputation is that the variability in standard errors over 1000 simulation runs is considerably higher for ten versus 100 imputed datasets when data are *missing at random*. Thus, although the benefit associated with creating and analysing higher numbers of imputed datasets is marginal for regression coefficients, doing so has a strong positive effect on the robustness of standard error estimates.

Standard errors for the VET Collection

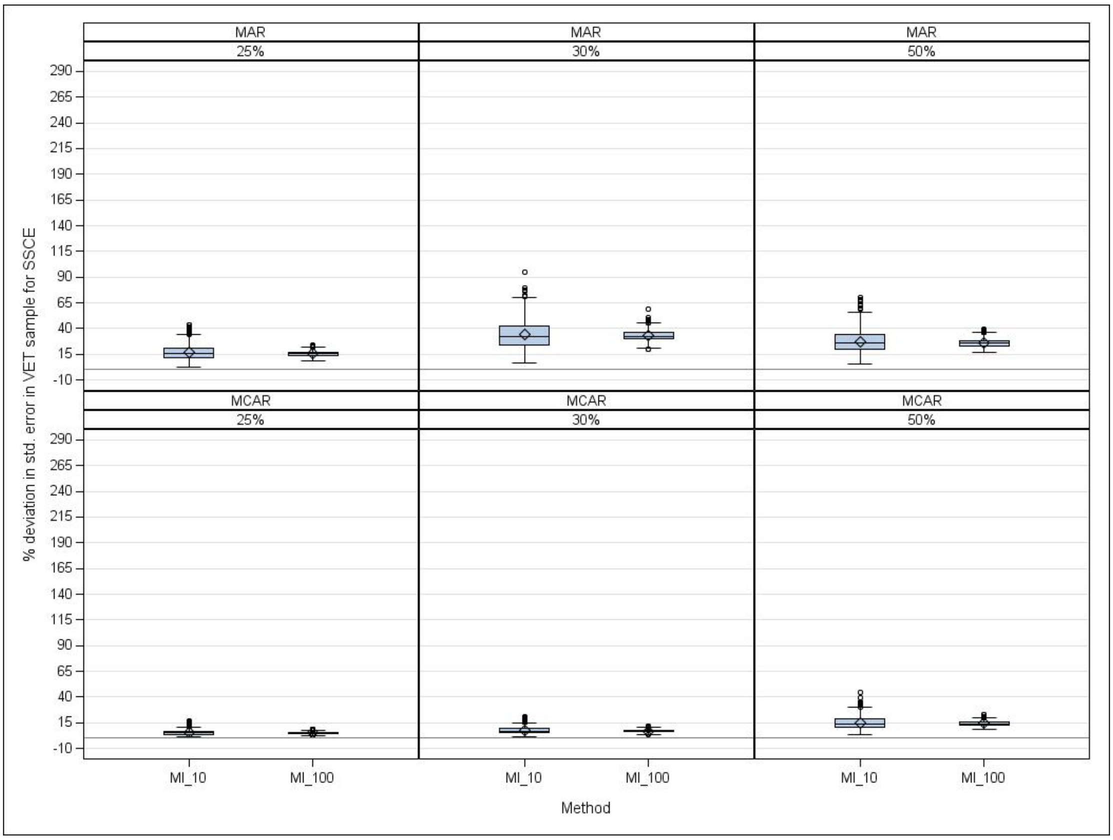
Recall that the main purpose of including the VET Collection in this study is to examine how different missing data methods handle missingness in binary predictors. Standard errors for the VET Collection are summarised in table 14.

Table 14 Percentage deviation in standard errors from the complete VET Collection sample

Predictor	Method	MCAR			MAR		
		25	30	50	25	30	50
Age	Listwise deletion	14.97	20.25	44.25	67.72	92.52	135.54
	Constant replacement	-1.44	-1.84	-3.17	-2.98	-2.69	-4.45
	Multiple imputation (10)	1.11	1.53	3.16	4.75	8.15	8.49
	Multiple imputation (100)	1.03	1.38	2.89	4.59	7.77	8.02
Disabled	Listwise deletion	15.16	20.39	46.43	51.48	65.15	727.85
	Constant replacement	-1.18	-1.55	-2.74	-0.37	1.89	-0.91
	Multiple imputation (10)	0.34	0.43	1.07	1.24	5.47	2.80
	Multiple imputation (100)	0.31	0.37	0.92	1.15	5.24	2.55
SSCE	Listwise deletion	14.97	20.21	44.14	96.51	110.47	191.64
	Constant replacement	1.45	2.11	5.02	8.85	17.97	13.04
	Multiple imputation (10)	5.47	7.45	15.09	16.43	33.96	27.44
	Multiple imputation (100)	5.11	6.92	14.39	15.44	32.88	25.98
Cert. III	Listwise deletion	15.03	20.15	44.23	50.47	71.30	105.81
	Constant replacement	1.62	2.28	5.41	13.76	16.77	21.41
	Multiple imputation (10)	5.39	7.29	14.80	25.23	28.81	38.60
	Multiple imputation (100)	5.03	6.81	13.78	24.60	27.38	36.37
NESB	Listwise deletion	15.01	20.22	44.31	71.37	182.22	147.10
	Constant replacement	2.81	3.54	7.93	9.77	48.47	23.16
	Multiple imputation (10)	5.45	6.93	13.52	14.15	49.10	29.88
	Multiple imputation (100)	4.80	6.02	12.18	12.82	48.01	28.96

Performance patterns for standard errors in the VET Collection were similar to those from LSAY for listwise deletion and constant replacement. A graphical representation of standard error results for the predictor *senior secondary certificate of education* is presented in figure 7. We note again that standard errors for the basic methods have been omitted from the figure, given that they are incorrect relative to those from the complete sample, based on theoretical grounds explained in conjunction with the LSAY sample above. Graphs for the standard errors of all other predictors in the VET Collection samples show a similar pattern and are provided in appendix E: Multiple imputation standard errors for LSAY and the VET Collection.

Figure 7 **Percentage deviation in standard errors from the complete VET Collection sample for** ***senior secondary certificate of education*** **using multiple imputation**



Note: LD = listwise deletion; CR = constant replacement; MI_10 = multiple imputation with 10 imputed datasets; MI_100 = multiple imputation with 100 imputed datasets.

Conclusion

In this paper, we have (1) provided key concepts around the handling of missing data in an easily accessible manner and (2) compared the performance of select missing data methods in two large-scale, mixed-variable social science datasets.

Our key points are as follows:

- Under realistic missing-data scenarios, listwise deletion and constant replacement perform poorly. In practical terms this means that both basic methods severely misjudge the impact of any single predictor variable on a given outcome of interest.
- Listwise deletion leads to much higher standard errors due to the loss of statistical power from discarding incomplete information. Constant replacement, on the other hand, underestimates standard errors. In practical terms this means that the mis-estimation of standard errors inherent in basic missing data methods greatly increases the potential for either failing to detect potentially important relationships between predictor variables and outcomes of interest, or for claiming the detection of such relationships where they do not exist. This, in turn, can lead to drawing wrong conclusions from research.
- Multiple imputation performs much better relative to the basic methods across all tested scenarios and samples. In practical terms this means that, even with high amounts of missing data, regression coefficients and standard errors remain stable and close to those of the complete-sample benchmarks.
- When using multiple imputation, increasing the number of imputed datasets results in more robust standard error estimates. In practical terms this means that using more imputations can increase the robustness of statistical analysis.

While modern missing data methods such as multiple imputation are no panacea for addressing every possible missing data scenario, they can often help reduce the risk of generating research results that are 'plain wrong'. We strongly encourage applied researchers to more carefully consider the potential impact of incomplete information and to use modern missing data methods whenever possible in their own analyses of large-scale social science collections.

References

Allison, PD 2002, *Missing data,* Sage, Thousand Oaks, California.

——2005, *Imputation of categorical variables with PROC MI*, viewed 30 August 2011, <http://www2.sas.com/proceedings/sugi30/113-30.pdf>

Azur, MJ, Stuart, EA, Frangakis, C & Leaf, PJ 2011, 'Multiple imputation by chained equations: what is it and how does it work?', *International Journal of Methods in Psychiatric Research*, vol.20, no.1, pp.40—9.

Enders, CK 2010, *Applied missing data analysis*, Guilford, New York.

Ganzeboom, HBG, de Graaf, PM, & Treiman, DJ 1992, 'A standard international socio-economic index of occupational status', *Social Science Research*, vol.21, no.1, pp.1—56.

Graham, JW 2007, *Missing data: analysis and design*, Institute of Education Sciences, viewed 30 August 2011, <http://ies.ed.gov/ncer/whatsnew/conferences/rct_traininginstitute/slides.asp?ppt=graham>.

Graham, JW, Olchowski, AE & Gilreath, TD 2007, 'How many imputations are really needed? Some practical clarifications of multiple imputation theory', *Preventative Science,* vol.8, no.3, pp.206—13.

Hershberger, SL & Fisher, DG 2003, 'A note on determining the number of imputations for missing data', *Structural Equation Modeling,* vol.10, no.4, pp.648—50.

Horton, N, Lipsitz, S & Parzen, M 2003, 'A potential for bias when rounding in multiple imputation', *The American Statistician*, vol.57, no.4, pp.229—33.

Lim, P 2011, *Weighting the LSAY cohorts*, NCVER, Adelaide.

Lipsitz, SR & Ibrahim, JG 1996, 'Using the EM algorithm for survival data with incomplete categorical covariates', *Lifetime Data Analysis,* vol.2, no.1, pp.5—14.

Little, RJA 1988, 'A test of missing completely at random for multivariate data with missing values', *Journal of the American Statistical Association,* vol.83, no.404, pp.1198—202.

——1992, 'Regression with missing X's: a review', *Journal of the American Statistical Association*, vol.87, no.420, pp.1227—37.

Little, RJA & Rubin, DB 2002, *Statistical analysis with missing data*, Wiley & Sons, Hoboken, New Jersey.

Marshall, A, Altman, DG, Royston, P & Holder, RL 2010, 'Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study', *BMC Medical Research Methodology*, vol.10, no.7, pp.1—16.

Mason, A, Best, N, Richardson, S & Plewis, I 2010, '*Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods'*, Imperial College, London.

Messer, K & Natarajan, L 2008, 'Maximum likelihood, multiple imputation and regression calibration for measurement error adjustment', *Statistics in Medicine*, vol.27, no.30, pp.6332—50.

Millsap, RE & Maydeu-Olivares, A 2009, *Quantitative methods in psychology*, Sage, Thousand Oaks, California.

Olinsky, A, Chen, S & Harlow, L 2003, 'The comparative efficacy of imputation methods for missing data in structural equation modeling', *European Journal of Operational Research*, vol.151, no.1, pp.53—79.

Piesse, A & Kalton, G 2009, *A strategy for handling missing data in the Longitudinal Study of Young People in England*, Department for Children, Schools and Families, viewed 30 August 2011, <https://www.education.gov.uk/publications/eOrderingDownload/DCSF-RW086.pdf>.

Rubin, DB 1976, 'Inference and missing data', *Biometrika*, vol.63, no.3, pp.581—92.

——1987, *Multiple imputation for nonresponse in surveys*, Wiley, New York.

Schafer, JL 1997, *Analysis of incomplete multivariate data*, CRC, Boca Raton, Florida.

Schafer, JL & Graham, JW 2002, 'Missing data: our view of the state of the art', *Psychological Methods*, vol.7, no.2, pp.147—77.

Schafer, JL & Olsen, MK 1998, 'Multiple imputation for multivariate missing data problems: a data analyst's perspective', *Multivariate Behavioral Research*, vol.33, no.4, pp.545—71.

Van Buuren, S & Groothuis-Oudshoorn, K 2009, 'MICE: multivariate imputation by chained equations in R', viewed 30 August 2011, <http://www.stefvanbuuren.nl/publications/MICE in R - Draft.pdf>.

White, IR, Royston, P & Wood, AM 2011, 'Multiple imputation using chained equations: issues and guidance for practice', *Statistics in Medicine*, vol.30, no.4, pp.377—99.

Yu, LM, Burton, A & Rivero-Arias, O 2007, 'Evaluation of software for multiple imputation of semi-continuous data', *Statistical Methods in Medical Research*, vol.16, no.3, pp.243—58.

Appendix A

Practical guidelines for applied researchers

In this appendix, we provide a few guidelines for applied social scientists who are faced with missing data problems. These guidelines are based on our practical experience with missing data in large-scale social science datasets. As such, the points we raise below are general in nature and by no means exhaustive. We differentiate between general guidelines and those specific to multiple imputation.

General guidelines

What should I know about my missing data?

It is crucial to familiarise yourself with the concrete missing data issue you are facing. Familiarising means understanding the 'why', 'how', and 'mechanism' of missingness, as discussed in the 'Behind the scenes' section. The 'why' and the 'how' will provide you with an indication of whether you are dealing with unit non-response, item non-response, wave non-response, attrition, or a mixture thereof. Remember that the methods we have outlined are suitable for item non-response. Unit non-response, wave non-response and attrition are better addressed through weighting (see Lim 2011 for weighting in LSAY; Piesse & Kalton 2009 for wave non-response weighting).

To determine whether your missing data are *completely* at random (the odds are that they are not), you should run a formal MCAR test. We suggest using Little's (1988) MCAR test because it is readily available in standard statistical software packages such as SPSS and SAS. The code to conduct the test in SAS is available online at <https://webapp4.asu.edu/directory/person/839490> (Enders 2011).

Remember that if the test rejects the MCAR hypothesis, there is no statistical test to ascertain whether your data are MAR or MNAR. If you have no further information on what caused the missingess, then MAR is the assumption of choice because both direct maximum likelihood and multiple imputation yield correct parameter estimates and standard errors under MAR. If you have additional information that provides a strong case for MNAR, then you can consider the Bayesian modelling techniques that have recently been proposed (see Mason et al. 2010). However, we caution readers that implementing these techniques requires considerable technical expertise in missing data methodology. A more practical option for less experienced researchers can therefore be to proceed with MAR-based methods (that is, maximum likelihood or multiple imputation), and being explicit about the associated bias in parameter estimates and standard errors.

Should I always use advanced missing data methods?

We do not generally recommend the use of constant replacement due to the resulting variance attenuation in the data. However, we suggest that listwise deletion is a feasible option where the amount of missing data is small (up to 5%, see also Schafer 1997). For all other scenarios we strongly encourage the use of multiple imputation or direct maximum likelihood.

Guidelines specific to multiple imputation

Which software should I use?

Given that the superiority of multiple imputation over basic missing data methods has been firmly established in the literature, multiple imputation functionality has become a standard option in all major statistical packages. Our practical experience is based on the use of multiple imputation in SAS

PROC MI and the Multiple Imputation by Chained Equations (MICE; Van Buuren & Groothuis-Oudshoorn 2009) package for R. For a detailed simulation study of diverse multiple imputation options in SAS, R, and Stata we refer the reader to Yu, Burton and Rivero-Arias (2007).

When comparing SAS PROC MI and R MICE, one potential shortcoming of the former is that it operates under a multivariate normal model, which means that no option is available to specify models for categorical or binary variables. However, 'in large datasets, with hundreds of variables of varying types, this is rarely appropriate' (Azur, Stuart & Frangakis 2011, p.41). The MICE package for R is more flexible because it allows the user to specify different imputation methods for different variable types (that is, predictive mean matching for continuous variables, multinomial logistic regression for categorical variables, and binary logistic regression for dichotomous variables). For this paper, we conducted some exploratory analyses comparing SAS and R for both our LSAY and VET Collection samples. Although results under R were closer to those obtained for the complete samples, differences between R and SAS were small. Overall, the Markov Chain Monte Carlo algorithm in SAS seems quite robust to departures from multivariate normality, producing good imputation results when variables are non-continuous.

Which variables should I include in my imputation model?

The 'golden rule' in specifying the imputation model is to include all predictor and outcome variables that are in the substantive analysis model. For instance, our substantive analysis model for the LSAY sample sought to predict completion of the senior secondary certificate of education from a respondent's sex, occupational aspirations, socioeconomic status, and mathematics achievement. Therefore, we included all four predictors and the outcome variable from the substantive analysis in the imputation model. If the imputed data are used for more than one substantive analysis, then the imputation model should contain every variable included in any of the analysis models.

Besides including all predictor and outcome variables of the substantive analysis model, more experienced researchers may also consider including additional variables that are not part of the substantive analysis model but which are predictors of the incomplete variables. This approach has been found to enhance the accuracy of the imputed data. The interested reader can find an in-depth treatise of specifying imputation models in White, Royston and Wood (2011).

How many imputed datasets should I create?

Between five and ten imputed datasets are generally considered sufficient for multiple imputation to perform well. When dealing with small samples or large amounts of missing data, up to 100 cycles have been suggested (see Graham, Olchowski & Gilreath 2007; Hershberger &Fisher 2003). Yet another study suggests a rule of thumb, whereby the number of imputations should be at least equal to the percentage of incomplete cases (White, Royston & Wood 2011).

For regression coefficients, our study indicated only a marginal performance benefit of creating 100 imputations over creating ten, even with 50% missingness. However, we ascertained a clear benefit of increasing the number of imputations with respect to the robustness of standard error estimates over 1000 runs. Since with modern statistical software packages the cost of creating a large number of imputations is virtually zero, we recommend using more rather than fewer imputed datasets.

Should I round the multiply imputed values?

SAS PROC MI provides optional settings that will round imputed values for categorical and binary variables to the nearest integer rather than return fractional imputed values. Rounding can be advantageous because we generally prefer imputed values to be plausible. For instance, if *sex* is coded 0 for males and 1 for females, the multiple imputation procedure should return a value of either 0 or 1 rather than a fractional value of 0.8. However, rounding in multiple imputation has been shown to produce bias with high levels of missing data (see Allison 2005; Horton et al. 2003).

We ran multiple imputation with and without the rounding option and found that rounding had a detrimental effect on results. Without rounding, results were notably closer to the 'true' results from the complete samples. The negative impact of rounding was particularly strong with the VET Collection sample, suggesting that rounding becomes more problematic as the number of incomplete binary variables increases. Therefore, we suggest that researchers refrain from rounding categorical and binary variables unless there is a compelling reason for limiting imputations to plausible values.

Is multiple imputation the perfect solution to missing data?

Absolutely not! The perfect solution to missing data is to avoid them in the first place. Short of this, every estimate of unobserved values is imperfect. Consequently, we do not argue in favour of using modern methods (that is, multiple imputation or direct maximum likelihood estimation) because we think they are perfect remedies. We argue in their favour because they are far better than anything else out there. Thus, in terms of producing sound social science research, modern missing data methods help us do the best we can in an imperfect world.

Appendix B

A worked example of multiple imputation

Researchers who are unfamiliar with multiple imputation sometimes consider the method a 'black box' that relies on dubious voodoo magic to make up data. To dispel this myth, we provide a simple worked example of how multiple imputation works in SAS. Our worked example is adapted from Enders (2010).

Suppose we have data on five respondents on a survey. Our data consists of respondents' (1) occupational aspirations scores at age 15 measured on a scale from 1 to 100, and (2) annual incomes at age 30. Suppose further that some of the occupational aspirations scores are missing, as shown in table B1.

Table B1 Example of dataset with partially missing occupational aspirations scores

Respondent	Occupational aspirations	Income ($)
1	.	65 000
2	.	50 000
3	17	40 000
4	24	55 000
5	98	75 000

Note: '.' indicates a missing value.

To perform multiple imputation on this dataset using the data augmentation algorithm, SAS would perform an imputation step and a posterior step.

The imputation step

We first use the available data to determine a regression equation that predicts occupational aspirations scores based on income. In our example, only cases three to five would be used to generate the regression equation. Suppose we use the following regression equation for the *i*th respondent

$$\text{OCC\_ASP}_i = -90 + 0.002 * \text{INCOME}_i.$$

We then add an error term to the predicted values from the regression model. The new equation for the *i*th respondent is,

$$\text{OCC\_ASP}_i = [-90 + 0.002 * \text{INCOME}_i] + z_i,$$

where z_i is a normally distributed error term with a mean of zero and a variance equal to the variance of the residuals from the regression of occupational aspirations on income. Note that z is different for each person in the dataset. (A different error term is drawn for each person.)

Suppose the variance $\hat{\sigma}^2{}_{OCC\_ASP|INCOME}$ = 7.2. The z terms would thus be drawn from a normal distribution with a mean of 0 and a variance of 7.2. The higher the amount of missing data, the higher we expect the variance to be.

We now impute the missing occupational aspirations scores for respondents 1 and 2 using the above regression model. The third column in table B2 lists the predicted occupational aspirations scores for respondents 1 and 2, the fourth column lists the random error terms (in this case z_1 = 2 .6 and z_2 = -

5.1), and the fifth column lists the imputed value, which is equal to the predicted value plus the random error term.

Table B2 Imputation results from the first iteration

Respondent	Occupational aspirations	Predicted value	Random error term (z_i)	Imputed value	Income ($)
1	.	40	2.6	***42.6***	65 000
2	.	10	-5.1	***4.9***	50 000

The imputed occupational aspirations scores are now passed on to the posterior step.

The posterior step

The first 'filled-in' dataset with imputed values from the previous step is illustrated in table B3.

Table B3 Filled-in dataset with imputation results from the first iteration

Respondent	Occupational aspirations	Income ($)
1	***42.6***	65 000
2	***4.9***	50 000
3	17	40 000
4	24	55 000
5	98	75 000

We now calculate the mean vector and covariance matrix of the first filled-in dataset. Suppose that

$$\hat{\mu}_{OCC\_ASP} = 37.3;\ \hat{\sigma}^2{}_{OCC\_ASP} = 900.$$

We now randomly perturb these mean and variance estimates. This is done by adding a random error term to $\hat{\mu}$ and $\hat{\sigma}^2$. Since the sampling distribution of the mean is a normal distribution with a standard deviation of $\frac{\hat{\sigma}}{\sqrt{N}}$, a new sample of five occupational aspiration scores should produce a mean that deviates from the current estimate by an average of $\frac{30}{\sqrt{5}} = 13.4$ score points. We thus generate a random error term from a normal distribution with a mean of 0 and a standard deviation of 13.4 and add it to $\hat{\mu}$. A similar process is used to generate a new variance estimate, but a different residual distribution is needed.

We now pass the perturbed mean and variance estimates back to the imputation step, which recalculates the regression equations using these *new* estimates of the means and variances. Repeating the imputation and posterior steps a large number of times creates multiple slightly different versions of the filled-in dataset. The question is: which version should we select as our first imputed dataset? This is where Markov Chain Monte Carlo theory comes in.

Markov Chain Monte Carlo

Suppose we want to take a random draw from a distribution of interest which has a complicated probability function. The idea behind using Markov Chain Monte Carlo for missing data is to take a 'random walk' that eventually settles on the distribution of interest. In this case, the distribution of interest is the distribution of all possible sets of replacement values for the missing data. Alternating between the imputation and the posterior step generates a series of estimates and corresponding 'filled-in' datasets:

$$Y_1^*,\ \theta_1^*,\ Y_2^*,\ \theta_2^*,\ Y_3^*,\ \theta_3^*, . . .,\ Y_t^*,\ \theta_t^*,$$

where Y_t* represents the imputed values from the imputation step and θ_t* contains the parameter estimates from the posterior step. This is like our ‘random walk’, since the imputation and posterior steps incorporate elements of randomness. If we create a long enough chain, we will eventually settle on the ‘stationary distribution’ of the chain, which is what we want to sample from.

We need to allow enough time for the Markov chain to ‘settle down’. This is called the ‘burn-in’ time. In SAS, the default number of burn-in iterations before each imputation is 200. After 200 iterations, we take the imputed values from the 201st iteration as our first imputed dataset. To ensure that the second imputed dataset is independent from the first, the chain is allowed to run for a further 100 iterations. We thus take the imputed values from the 301st iteration as our second imputed dataset. SAS allows 100 iterations between each imputed dataset.

A note on imputation with more than one incomplete variable

When the dataset has more than one variable, the regression stage of the imputation step is slightly more complicated. Suppose we have three variables, X_1, X_2, X_3, which all have some amount of missing data. This results in having six possible missing data scenarios, whereby we could have cases with missing data on (1) X_1 only, (2) X_2 only, (3) X_3 only, (4) X_1 and X_2, (5) X_1 and X_3, or (6) X_2 and X_3. In the imputation step, we would need a unique regression equation for each scenario, as well as a different residual term for each error distribution that generates the z_i terms. While this seems complex in theory, these steps are performed automatically by the statistical software package.

Appendix C

Further details of the MCAR and MAR mechanisms

Choosing 'sensible' logit coefficients for imposing MAR

Here, we use our LSAY sample to illustrate how we chose 'sensible' coefficients for our implementation of the MAR mechanism. Remember that in our example we would like the probability of X_3 (a respondent's mathematics achievement score) being missing to depend on

- Y (WHETHER OR NOT THE PERSON COMPLETED SCHOOL)
- $X1$ (THE PERSON'S OCCUPATIONAL ASPIRATION SCORE)
- $X2$ (THE PERSON'S SOCIOECONOMIC STATUS)

and for this probability to *increase* if X_1 is already missing. These rules can be modelled with the following logit function,

$$\text{logit}[\Pr_{(\text{MATHS MISSING})}] = \alpha_0 + \alpha_1 Y + \alpha_2 X_1 + \alpha_3 X_3 + \alpha_4 MX_1,$$

where α_0, α_1, α_2, α_3, and α_4 are constants, and MX_1 is a missingness indicator for predictor X_1, which equals 1 if the value of X_1 is missing, and 0 if it is observed. We will refer to the value of logit[$\Pr_{(\text{MATHS MISSING})}$] as the 'z-score' for an observation.

How did we choose values for α_0, α_1, α_2, α_3, and α_4? The main goal governing our choice of logit coefficients was to achieve the desired overall level of missingness for each of our samples (that is, 25%, 50% etc.), but our simultaneous goal was to ensure that the coefficients were sensible.

We first provide some intuition as to what the coefficients mean. The intercept α_0 can be interpreted as the 'baseline probability' for mathematics achievement being missing for an average person (that is, someone who has a score of 0 across all standardised predictors).

In our context, an average person is someone who *did* complete the senior secondary certificate of education (Y = 0), who has an *average* occupational aspirations score (X_1 = 0), and an *average* socioeconomic status score (X_2 = 0). We thus set the probability of deletion for an average person to be fairly low. For example, if we let $\alpha_0 = -1.38$, then $\Pr_{(\text{MATHS MISSING})} = \frac{1}{1+e^{-1.38}} = .20$ (using the inverse logit function). That is, an average person has a 20% probability of having their mathematics achievement score missing. This probability increases or decreases as we change the values of the other variables in the dataset.

When the values of Y, X_1 and X_2 change, we want the probability of deletion to change accordingly. For example, if a respondent *did not* complete the senior secondary certificate of education (Y = 1), we want that respondent's probability of deletion to *increase*. The z-score thus needs to increase, and the coefficient attached to Y (that is, α_1) should be positive. Also, if MX_1 = 1 (that is, if X_1 is already missing for that respondent), we want the probability of deletion for X_3 to increase further. Consequently, α_4 needs to be positive.

If a respondent has an *above average* occupational aspirations score, we want the probability of deletion to *decrease*; thus the coefficient attached to X_1 should be negative. Similarly, if the respondent has an *above average* socioeconomic status score, we want the probability of deletion to decrease. The coefficient attached to X_2 should then also be negative.

In general, if we want a positive value of a variable to *increase* the probability of a value being missing, we make the coefficient associated with that variable in the logit function *positive*. Likewise, if we want a positive value of a variable to *decrease* the probability of a value being missing, we make the coefficient associated with that variable *negative*.

Once the coefficients of the logit functions are set, the remainder of the deletion process for MAR is very similar to that for MCAR. For each case, we first calculate its z-score using the logit function. We then find the corresponding probability of deletion using the inverse logit function. For example, if logit($\mathrm{Pr}_{(\text{MATHS MISSING})}$) = -0.37 (the z-score), then $\mathrm{Pr}_{(\text{MATHS MISSING})} = \frac{1}{1+e^{-0.37}}$ = .41. Next, we generate a uniform random variable, *u*, from the [0,1] distribution. The predictor value for a case is set to missing if *u* is less than the calculated probability. Notice that in contrast to MCAR, the deletion probabilities for MAR are *not* the same for every respondent (see table C1).

Table C1 Illustration of data deletion process for MAR

Person	z-score	$\mathrm{Pr}_{(\text{MATHS MISSING})}$	*u* ~ [0,1]	Delete MATHS if *u < Pr*
1	-.37	.41	.315	Delete
2	2.19	.10	.224	Keep
3	-2.42	.08	.146	Keep

We repeat this process for each predictor that needs to be made missing, and adjust all of the coefficient values to achieve the desired overall level of missingness. The final coefficients used are given in the next section.

Actual MAR and MCAR mechanisms used

In this section we give (i) the actual coefficients for the logit functions used to impose our MAR mechanisms, and (ii) the deletion probabilities used for our MCAR mechanisms. We first specify these mechanisms for the LSAY dataset, followed by the VET dataset.

LSAY dataset

In the following tables, the variables X_0 to $X_{3,}$ Y, MX_1 and MX_2 refer to the following:

X_0 = SEX, X_1 = OCC. ASP, X_2 = SES, X_3 = MATH

Y = SSCE COMPLETION STATUS,

MX_1 = indicator variable which is 1 if X_1 is missing, 0 if X_1 is not missing

MX_3 = indicator variable which is 1 if X_3 is missing, 0 if X_3 is not missing

The logit functions for the MAR mechanisms are given in tables C2, C3 and C4. Each logit function determines the probability of deletion of a particular variable. We present the logit functions in tabular form: reading across a row from left to right gives the full logit function for one variable. A shaded cell indicates that the variable is not allowed to be part of the function. For instance, in table C2, the probability that X_0 is missing cannot depend on X_0 itself, therefore X_0's coefficient is greyed out for X_0's logit function.

Table C2 Logit functions for MAR with 50% missingness

Parameter	Intcpt	Y	X_0	X_1	X_2	X_3	MX_1	MX_3
Logit(Pr (X_0) missing) =	-1.8	+1(Y)		-0.8(X_1)	-0.5(X_2)			
Logit(Pr (X_1) missing) =	-1.8	+0.5(Y)			-0.5(X_2)	-0.5(X_3)		
Logit(Pr (X_2) missing) =	-2	+0.7(Y)		-0.6(X_1)		-0.6(X_3)		+1(MX_3)
Logit(Pr (X_3) missing) =	-1.9	+0.5(Y)			-0.6(X_2)		+1(MX_1)	

Read across a row of table C2 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{2\ \text{MISSING}})] = -2 + 0.7(Y) + (-0.6)(X_1) + (-0.6)(X_3) + 1(MX_3).$$

Table C3 Logit functions for MAR with 25% missingness

Parameter	Intcpt	Y	X_0	X_1	X_2	X_3	MX_1	MX_3
Logit(Pr (X_0) missing) =	-2.9	+1(Y)		-0.5(X_1)	-0.5(X_2)			
Logit(Pr (X_1) missing) =	-2.9	+1(Y)			-0.4(X_2)	-0.5(X_3)		
Logit(Pr (X_2) missing) =	-3	+1(Y)		-0.3(X_1)		-0.3(X_3)		+2(MX_3)
Logit(Pr (X_3) missing) =	-3	+1(Y)			-0.4(X_2)		+2(MX_1)	

Read across a row of table C3 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{0\ \text{MISSING}})] = -2.9 + 1(Y) + (-0.5)(X_1) + (-0.5)(X_2).$$

Table C4 Logit functions for MAR with 17% missingness

Parameter	Intcpt	Y	X_0	X_1	X_2	X_3	MX_1	MX_3
Logit(Pr (X_1) missing) =	-2	+1(Y)			-0.4(X_2)	-0.4(X_3)		
Logit(Pr (X_2) missing) =	-6	+1(Y)		-0.5(X_1)		-0.5(X_3)	+1(MX_1)	

Read across a row of table C4 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{1\ \text{MISSING}})] = -2 + 1(Y) + (-0.4)(X_2) + (-0.4)(X_3).$$

The deletion probabilities for the three MCAR mechanisms are given in table C5. Note that they are constant within each level of missingness; this is to ensure that the total missingness is evenly distributed across the variables.

Table C5 Deletion probabilities for MCAR with 25%, 50% and 17% missingness

25% Missingness	50% Missingness	17% Missingness
Pr (X_0 missing) = 0.07	Pr (X_0 missing) = 0.15	Pr (X_0 missing) = 0.05
Pr (X_1 missing) = 0.07	Pr (X_1 missing) = 0.15	Pr (X_1 missing) = 0.05
Pr (X_2 missing) = 0.07	Pr (X_2 missing) = 0.15	Pr (X_2 missing) = 0.05
Pr (X_3 missing) = 0.07	Pr (X_3 missing) = 0.15	Pr (X_3 missing) = 0.05

VET Collection

In the following tables, the variables X_1 to X_3, Y, MX_1 and MX_2 refer to the following:

X_1= SEX, X_2 = CERT. III, X_3 = NESB

Y = DROPOUT

MX_1 = indicator variable which is 1 if X_1 is missing, 0 if X_1 is not missing

MX_2 = indicator variable which is 1 if X_2 is missing, 0 if X_2 is not missing

The logit functions for the MAR mechanisms are given in tables C6, C7 and C8. We present the logit functions in tabular form: reading across a row from left to right gives the full logit function for one variable. A shaded cell indicates that the variable is not allowed to be part of the function. For instance, in table C6, the probability that X_1 is missing cannot depend on X_1 itself, therefore X_1's coefficient is greyed out for X_1's logit function.

Table C6 Logit functions for MAR with 50% missingness

Parameter	Intcpt	Y	X_1	X_2	X_3	MX_1	MX_2
Logit(Pr (X_1) missing) =	-1.5	+1.3(Y)		-2(X_2)	+1(X_3)		
Logit(Pr (X_2) missing) =	-1.8	+1.2(Y)	+1(X_1)		+0.5(X_3)	+0.4(MX_1)	
Logit(Pr (X_3) missing) =	-1.6	+1.5(Y)	+1.7(X_1)	-1.5(X_2)			

Read across a row of table C6 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{2\ \text{MISSING}})] = -1.8 + 1.2(Y) + 1(X_2) + 0.5(X_3) + 0.4(MX_1).$$

Table C7 Logit functions for MAR with 25% missingness

Parameter	Intcpt	Y	X_1	X_2	X_3	MX_1	MX_2
Logit(Pr (X_1) missing) =	-2.7	+2(Y)		-2(X_2)	+0.8(X_3)		
Logit(Pr (X_2) missing) =	-3	+1.8(Y)	+1.2(X_1)		+0.5(X_3)	+1(MX_1)	
Logit(Pr (X_3) missing) =	-2.8	+1.5(Y)	+2(X_1)	-1.5(X_2)			

Read across a row of table C7 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{2\ \text{MISSING}})] = -3 + 1.8(Y) + 1.2(X_2) + 0.5(X_3) + 1(MX_1).$$

Table C8 Logit functions for MAR with 30% missingness

Parameter	Intcpt	Y	X_1	X_2	X_3	MX_1	MX_2
Logit(Pr (X_1) missing) =	-2.32	+2.2(Y)		-1(X_2)	+1.5(X_3)	0	0
Logit(Pr (X_2) missing) =	-4.58	+2.2(Y)	+2(X_1)		+2(X_3)	+2(MX_1)	0
Logit(Pr (X_3) missing) =	-2.75	+2.1(Y)	+1.2(X_1)	-1.2(X_2)		0	+2(MX_2)

Read across a row of table C8 to obtain the full logit function for that variable, for example,

$$\text{logit}[\Pr(X_{2\ \text{MISSING}})] = -4.58 + 2.2(Y) + 2(X_1) + 2(X_3) + 2(MX_1).$$

The deletion probabilities for the three MCAR mechanisms are given in table C9. Note that they are constant down each column; this is to ensure that the total missingness is evenly distributed across the variables.

Table C9 Deletion probabilities for MCAR with 25%, 50%, and 30% missingness

25% Missingness	50% Missingness	30% Missingness
Pr (X_0 missing) = 0.09	Pr (X_0 missing) = 0.21	Pr (X_0 missing) = 0.12
Pr (X_1 missing) = 0.09	Pr (X_1 missing) = 0.21	Pr (X_1 missing) = 0.12
Pr (X_2 missing) = 0.09	Pr (X_2 missing) = 0.21	Pr (X_2 missing) = 0.12

Percentage missingness per predictor

A final criterion in setting the logit coefficients and deletion probabilities was that the missing values should be evenly distributed across the predictors, to avoid all missing values being concentrated in one variable. Tables C10 and C11 illustrate the level of missingness imposed on the predictors for the LSAY and VET Collection samples, respectively, as a result of applying the above logit functions and deletion probabilities.

Table C10 Percentage missingness imposed on predictors in the LSAY sample

Mechanism	Sex	Occ. asp	SES	Maths
MAR with 25% missingness	8	8	10	10
MAR with 50% missingness	22	19	21	20
MAR with 17% missingness	0	16	1	0
MCAR with 25% missingness	7	7	7	7
MCAR with 50% missingness	15	15	15	15

Table C11 Percentage missingness imposed on predictors in the VET Collection sample

Mechanism	SSCE	Cert. III	NESB
MAR with 25% missingness	8	13	10
MAR with 50% missingness	21	25	21
MAR with 30% missingness	19.2	14.5	14.6
MCAR with 25% missingness	9	9	9
MCAR with 50% missingness	21	21	21

Notice that the missingness levels across each row of tables C10 and C11 are approximately constant, apart from MAR 17 in table C10 and MAR 30 in table C11, which were designed to mimic the observed percent missingness in the true dataset.

Appendix D

Regression coefficients for LSAY and the VET Collection

Figure D1 Percentage deviation in regression coefficients for *sex* (LSAY)

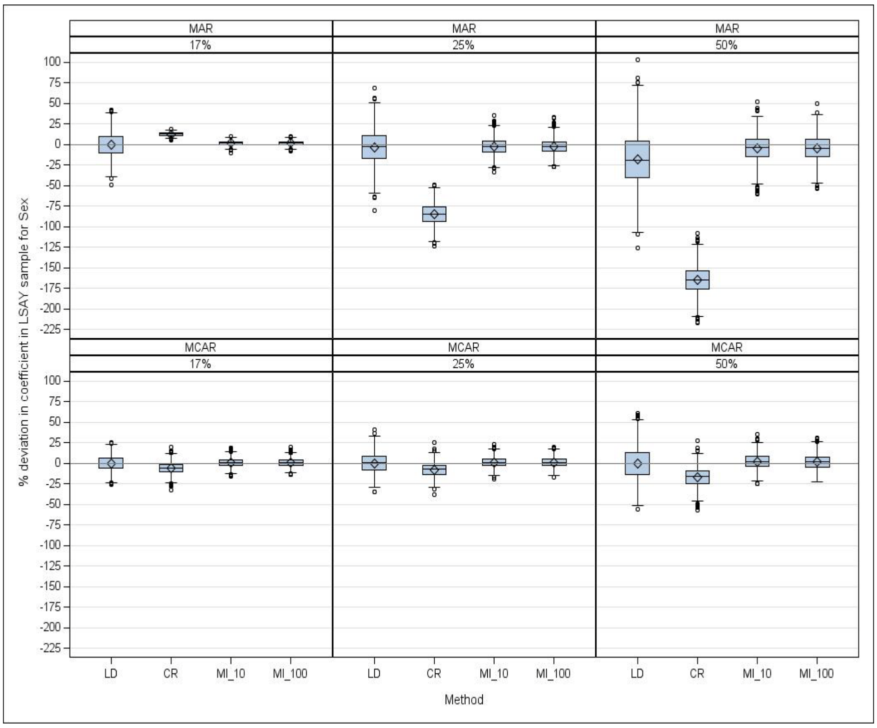


Figure D2 Percentage deviation in regression coefficients for *occupational aspirations* (LSAY)

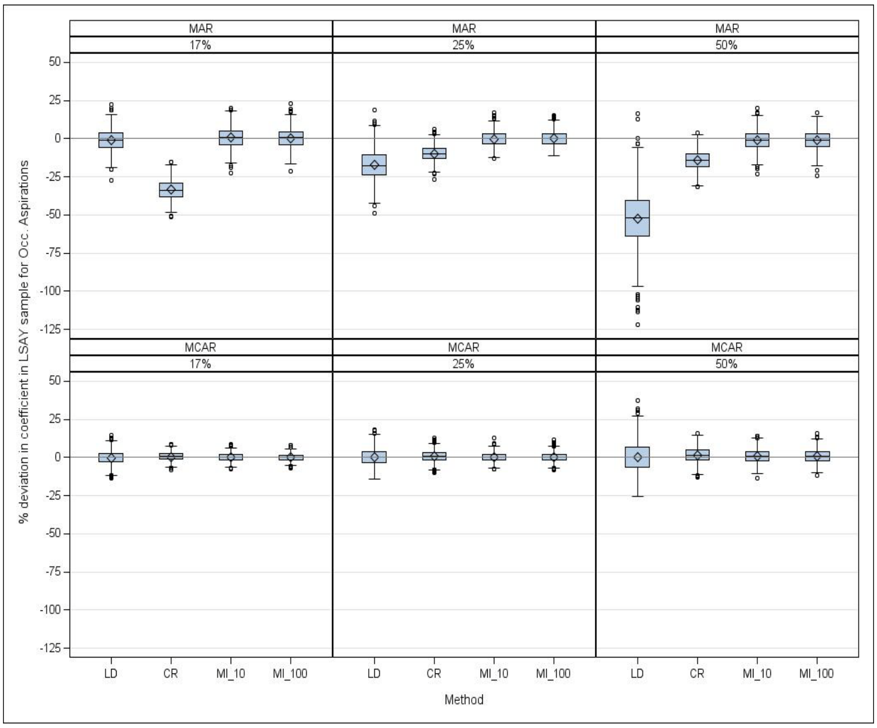


Figure D3 Percentage deviation in regression coefficients for *socioeconomic status* (LSAY)

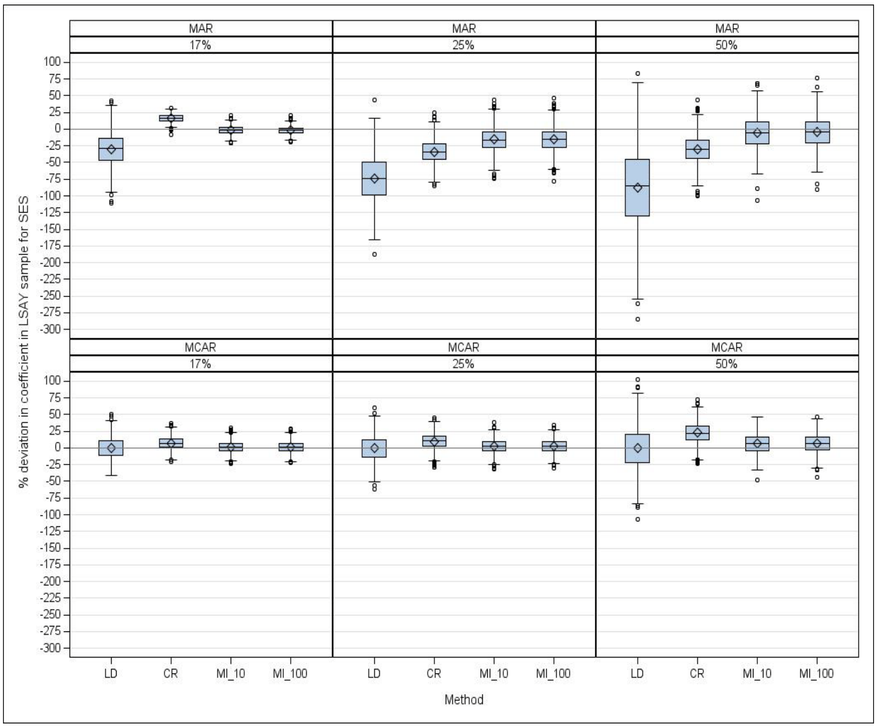


Figure D4 Percentage deviation in regression coefficients for *mathematics achievement* (LSAY)

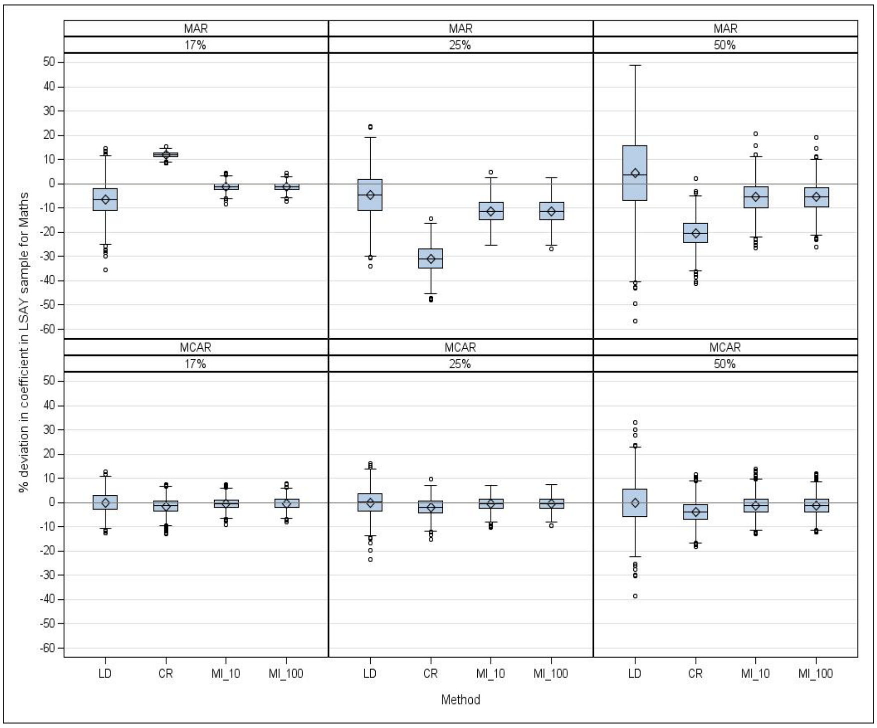


Figure D5 Percentage deviation in regression coefficients for *age* (VET Collection)

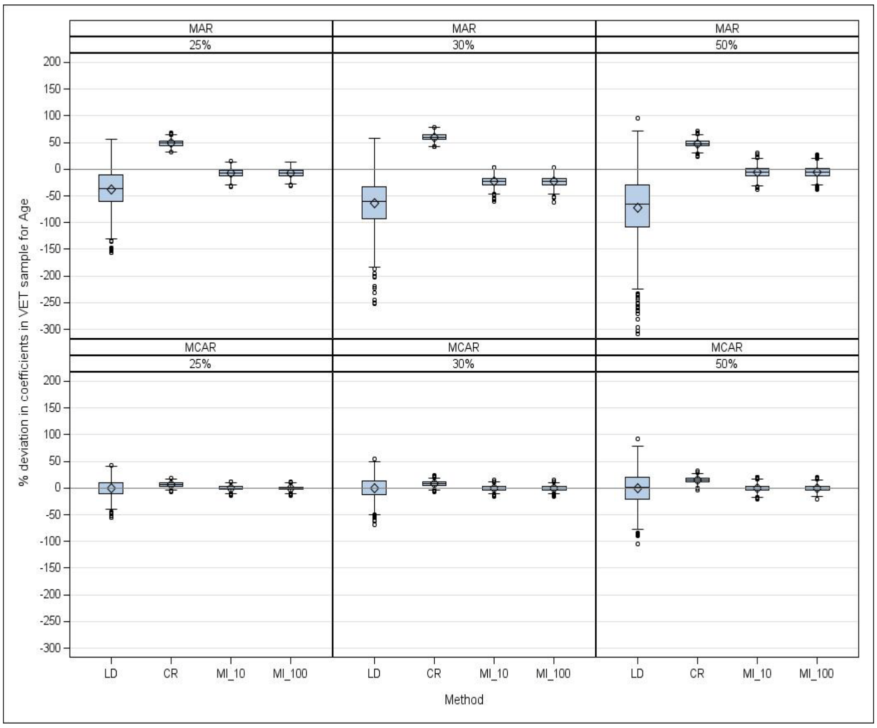


Figure D6 Percentage deviation in regression coefficients for *disabled* (VET Collection)

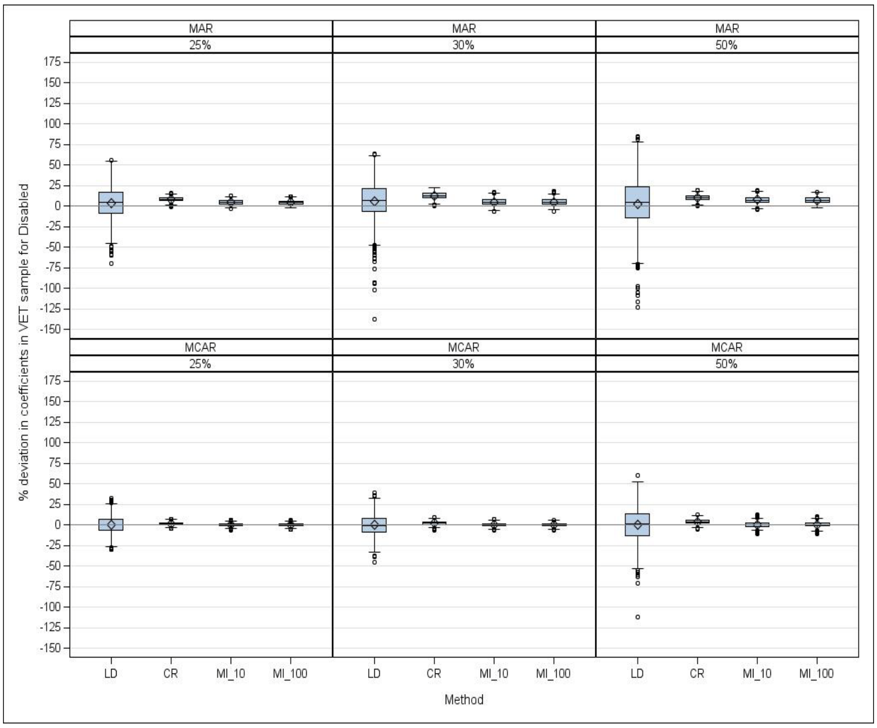


Figure D7 Percentage deviation in regression coefficients for *senior secondary certificate of education* (VET Collection)

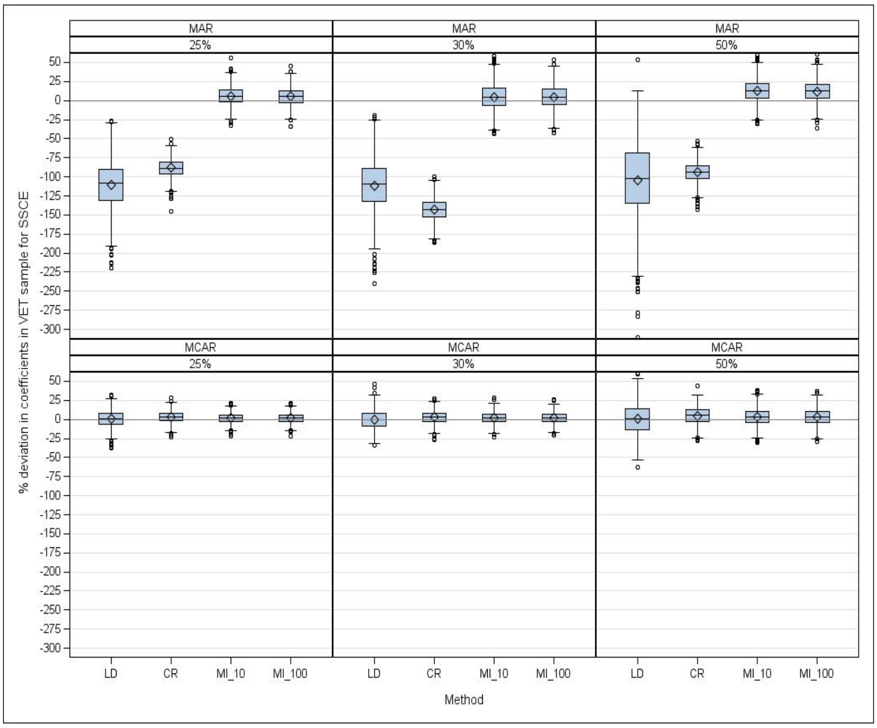


Figure D8 Percentage deviation in regression coefficients for *vocational certificate III* (VET Collection)

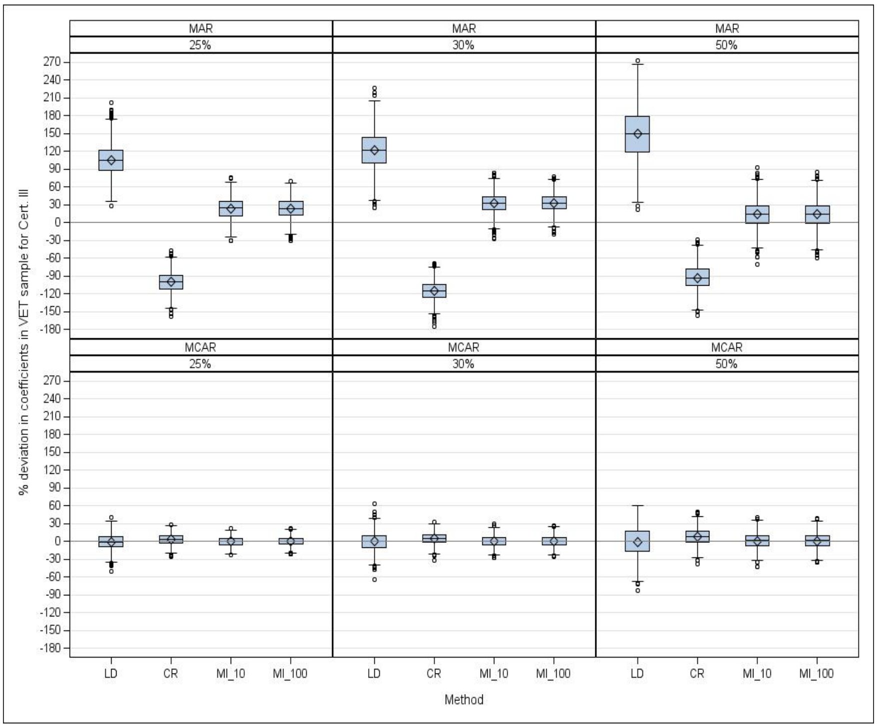
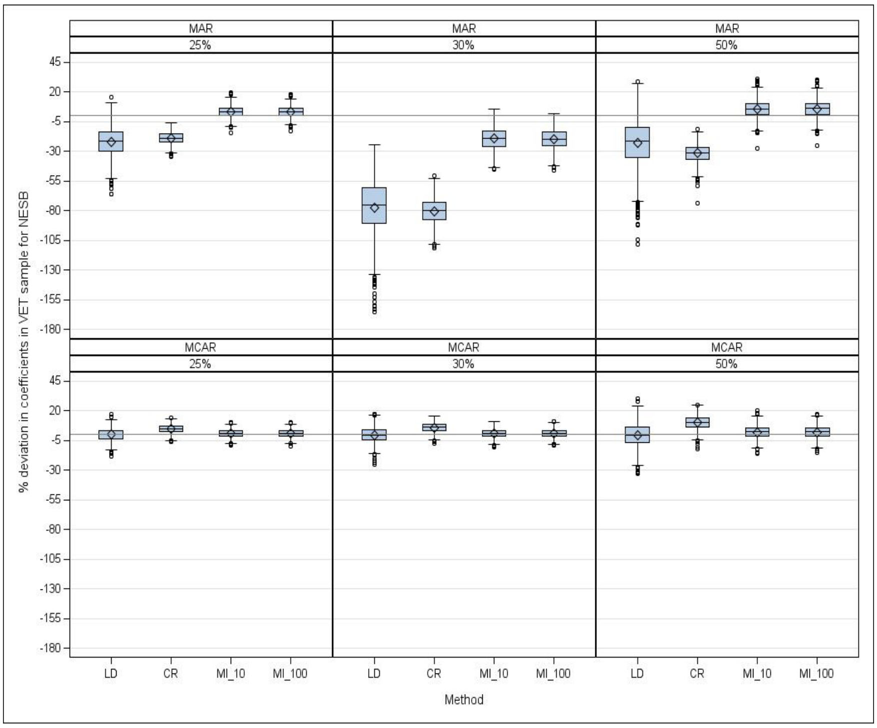


Figure D9 Percentage deviation in regression coefficients for *non-English speaking background* (VET Collection)



Appendix E

Multiple imputation standard errors for LSAY and the VET Collection

Figure E1 Percentage deviation in standard errors for *sex* (LSAY)

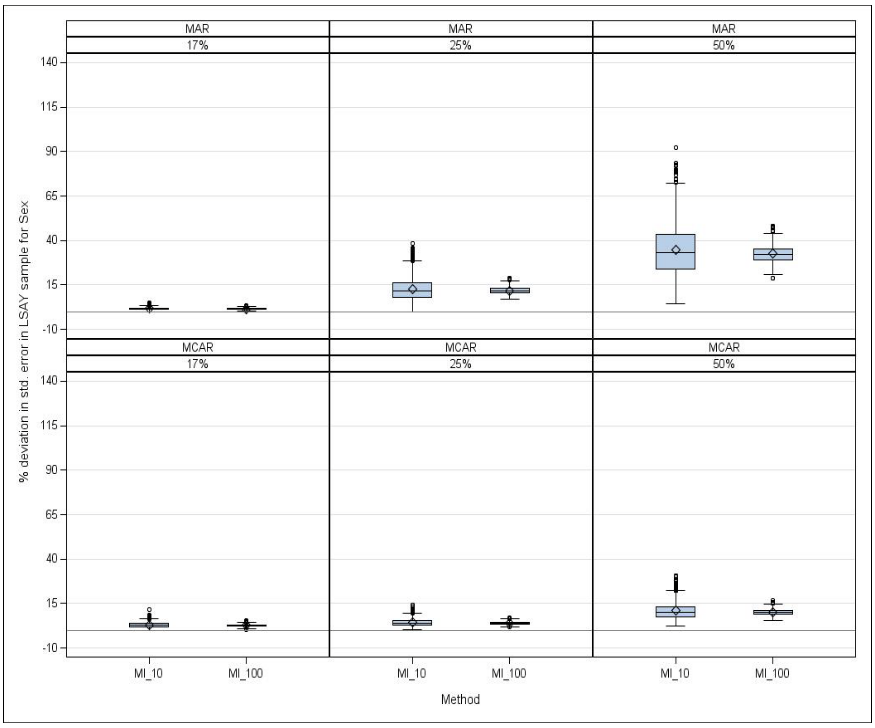


Figure E2 Percentage deviation in standard errors for *occupational aspirations* (LSAY)

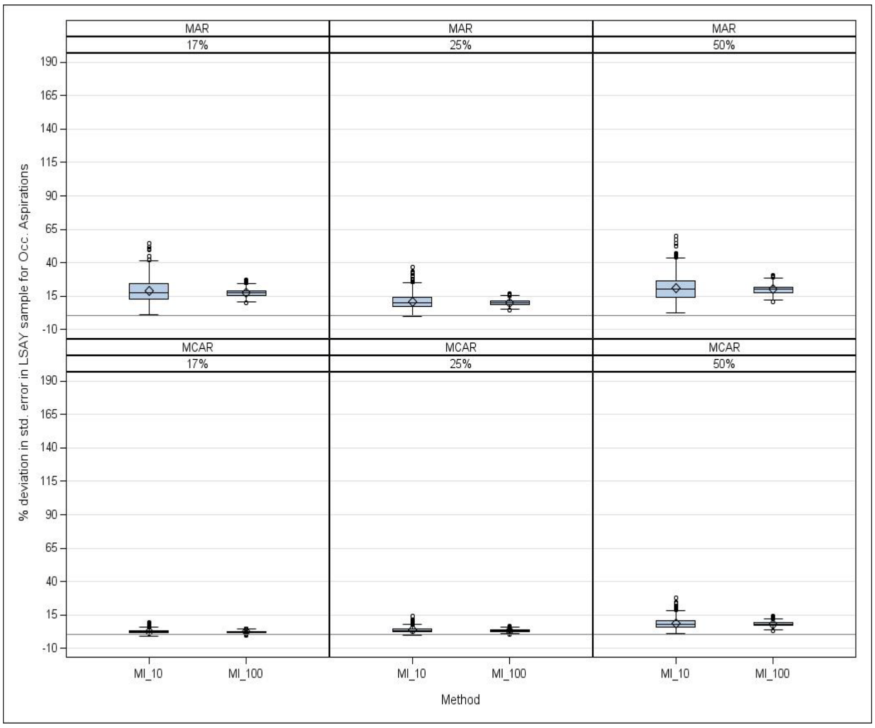


Figure E3 Percentage deviation in standard errors for *socioeconomic status* (LSAY)

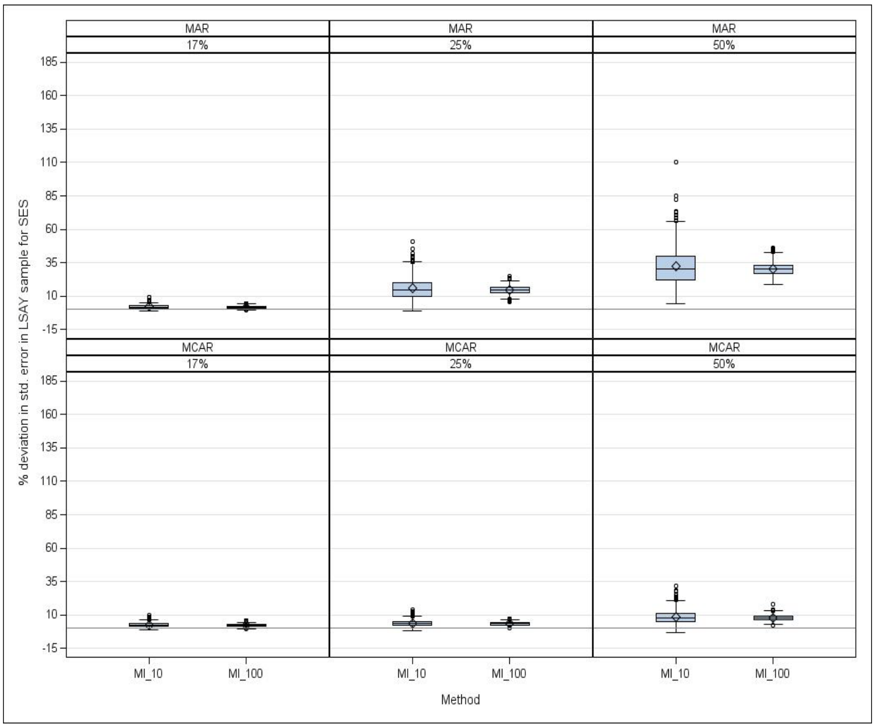


Figure E4 Percentage deviation in standard errors for *mathematics achievement* (LSAY)

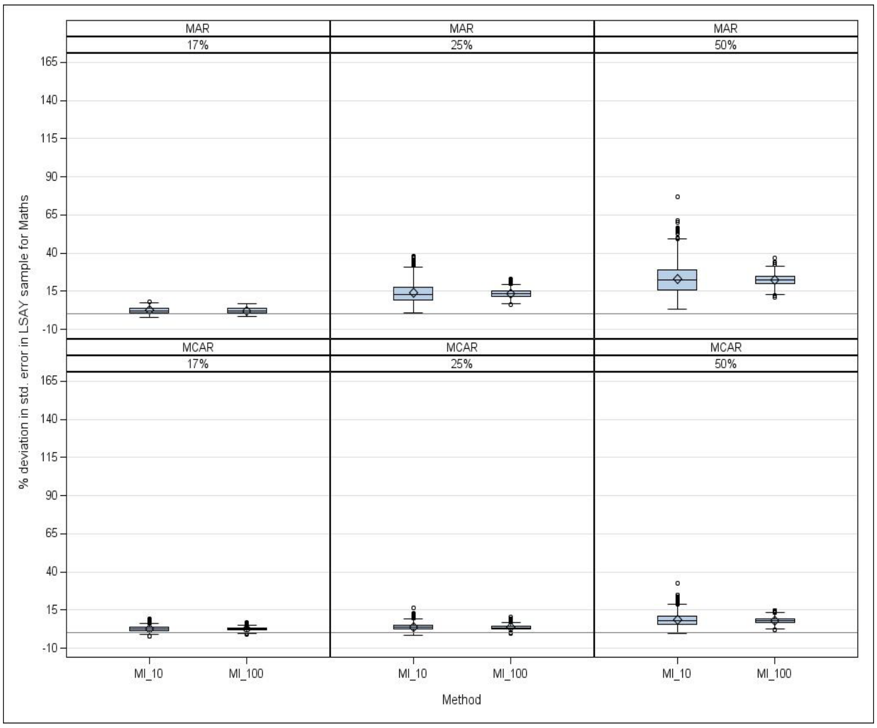


Figure E5 Percentage deviation in standard errors for *age* (VET Collection)

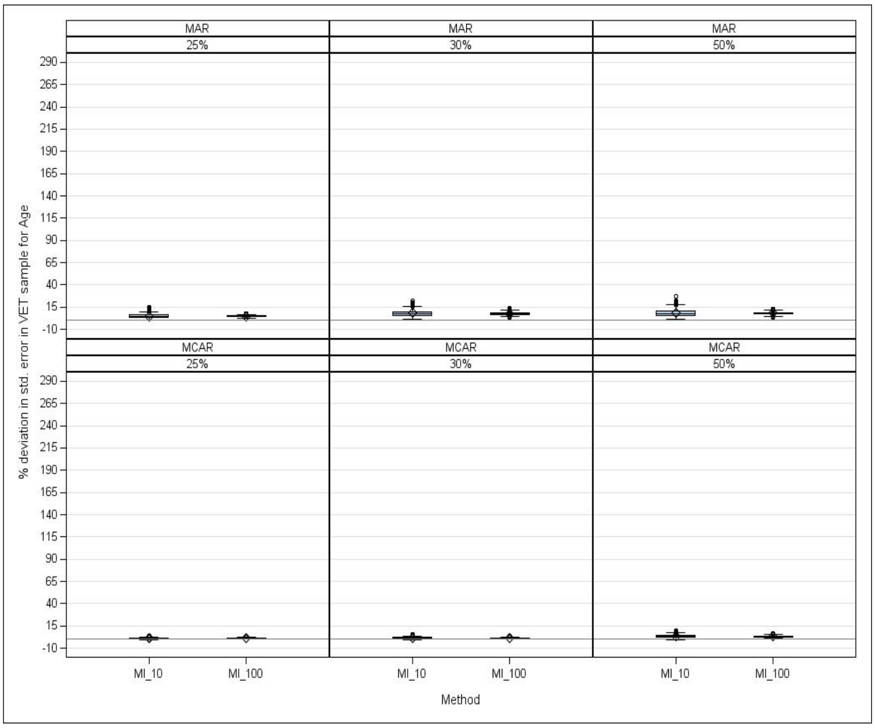


Figure E6 Percentage deviation in standard errors for *disabled* (VET Collection)

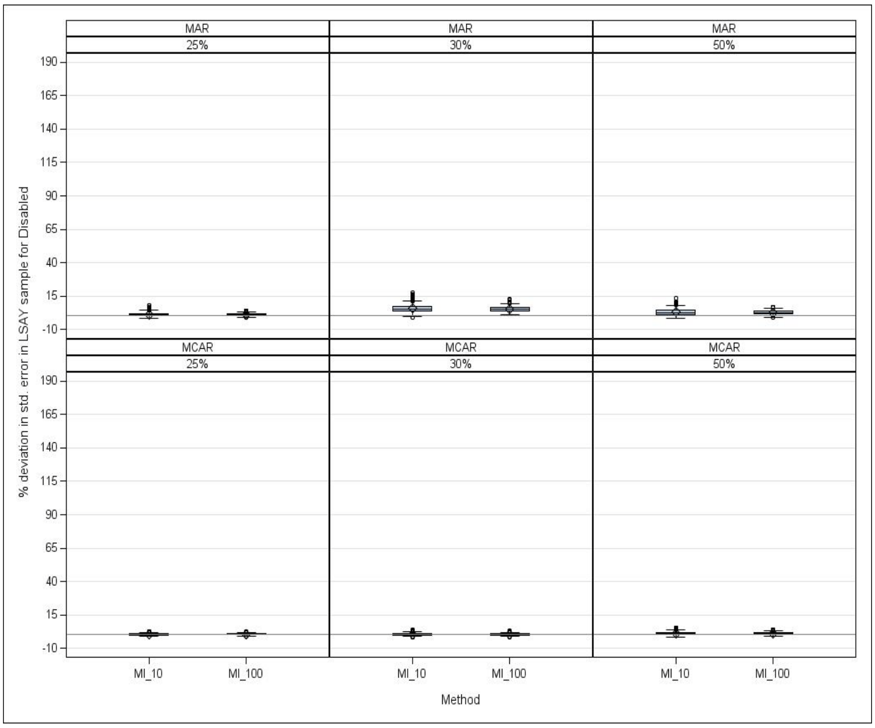


Figure E7 Percentage deviation in standard errors for *senior secondary certificate of education* (VET Collection)

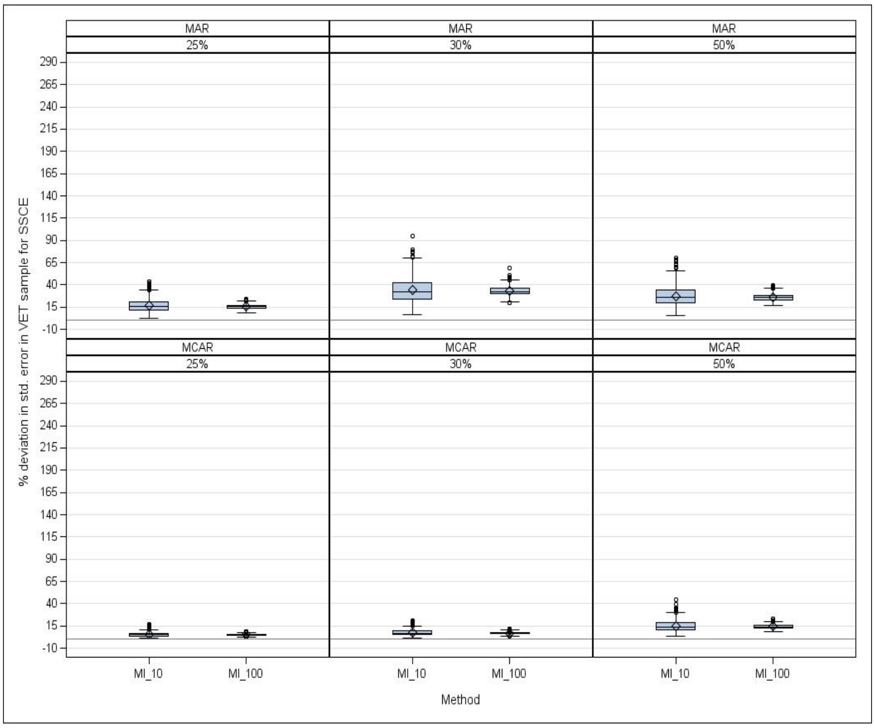


Figure E8 Percentage deviation in standard errors for *vocational certificate III* (VET Collection)

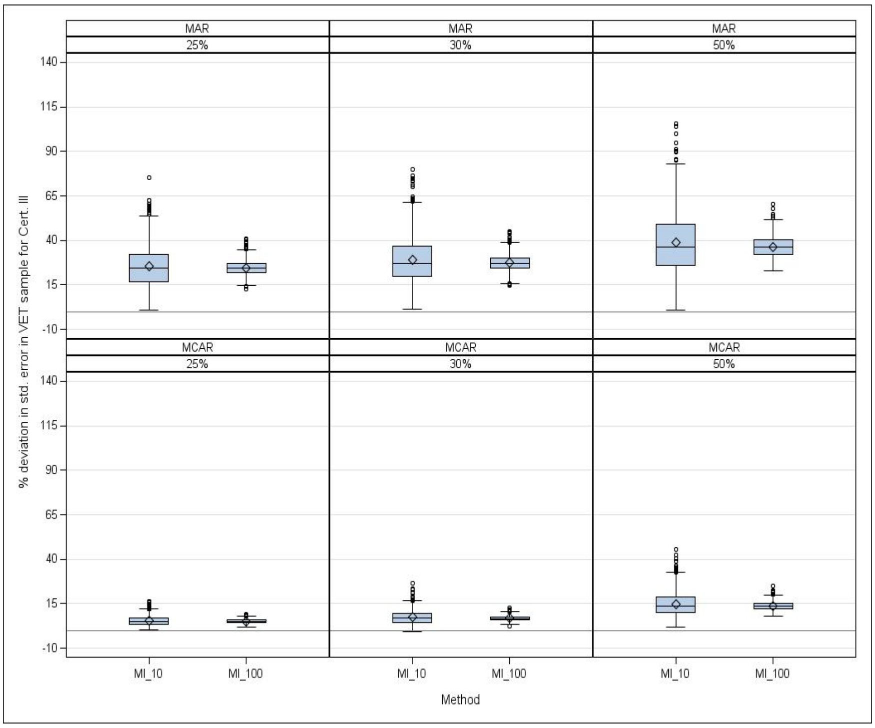
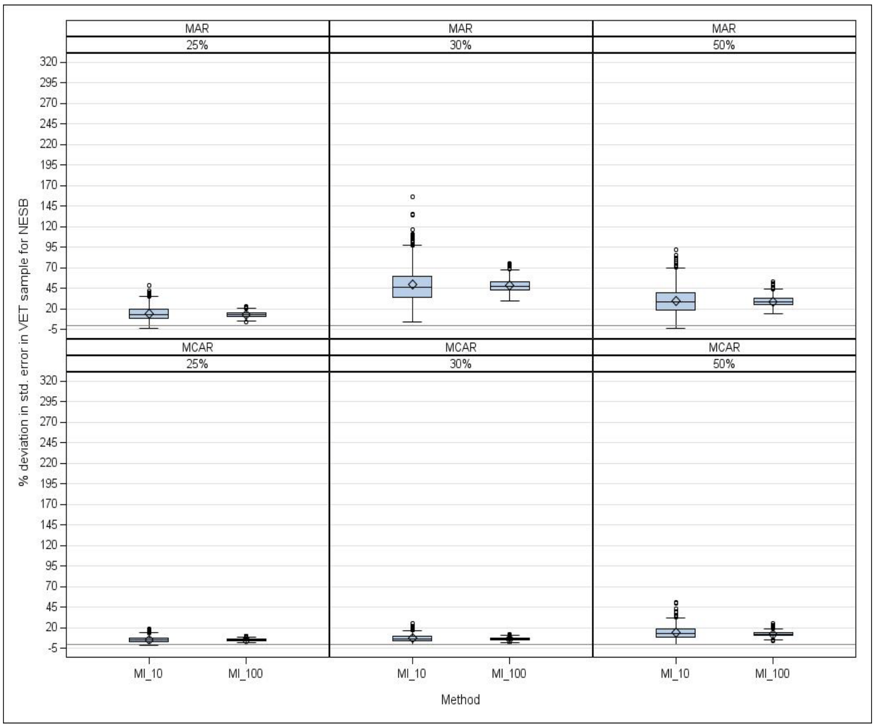


Figure E9 Percentage deviation in standard errors for *non-English speaking background* (VET Collection)



Appendix F

SAS code

Here we provide the basic SAS code for the different missing data methods included in our performance test. Interested readers should consult the SAS online documentation at <http://support.sas.com/documentation/92/index.html> for details on using SAS PROC MI and SAS PROC MIANALYZE.

Listwise deletion

SAS automatically applies listwise deletion when analysing a dataset with missing values. For example, if fitting a logistic regression model, any observation with one or more missing values will be automatically excluded from the analysis. We used the standard logistic regression routines in SAS to analyse our samples under listwise deletion. Below is an example for carrying out listwise deletion on an LSAY sample with 25% of data missing completely at random. Since listwise deletion is the default method, we actually just run the straight logistic regression analysis. However, since LSAY contains attrition weights we need to use PROC SURVEYLOGISTIC in SAS. The 'ods' output statement is used to save the parameter estimates in a file called "LD_all".

```
/* Save parameter estimates in a file called 'LD_all' */

ods output Surveylogistic.ParameterEstimates = LD_all;

proc surveylogistic data=work.mcar25_lsay;

        class sex (ref = '0')/ param=ref;

        model dropout = sex occ_asp escs math_std;

        weight weight;

run; ods output close;
```

Constant replacement

Constant replacement includes mean and mode substitution. For each variable, any missing values were replaced with either the mean or mode of that variable, depending on whether the variable was continuous or binary. For example, the mode of the predictor *sex* in the LSAY sample is 0; any missing values in *sex* were thus replaced with a value of 0. Similarly, the mean of *mathematics achievement* is 0 because that predictor was standardised. Any missing maths scores were therefore replaced with a value of 0. Below is an example for performing constant replacement on an LSAY dataset with 25% of values missing completely at random.

```
/* Create Constant Replacement dataset */
data matching.mcar25_lsay_CR_Dataset;

  set work.mcar25_lsay;

        if escs = . then

                escs = 0; *0 is the mean of escs;

        if occ_asp = . then

                occ_asp = 0; *0 is the mean of occ_asp;

        if math_std = . then
```

```
            math_std = 0; *0 is the mean of math_std;
    if sex = . then
      sex = 1; *1 is the "mode" of sex; run;
```

Multiple imputation

Multiple imputation was implemented using the PROC MI routine with both ten and 100 imputations. An example code is given below for the LSAY and VET Collection datasets, respectively. There are additional rounding and max/min options available in PROC MI, which forces the imputed values to fall within the range of the observed values. However, in our analysis we did not use these options. The arguments for and against rounding the imputed values are presented in appendix A: Practical guidelines for applied researchers.

LSAY

(has weights, uses PROC SURVEYLOGISTIC, which requires using the “parms=” option in PROC MI ANALYZE).

```
*/ PROC MI will create the specified number of imputed datasets */;

*/ In this case we have set the no. of imputations to 10 */;

proc mi data = work.mcar25_lsay nimpute= 10 seed=12345 out=imputed;

var sex occ_asp escs math_std dropout;

run;

*/ PROC SURVEYLOGISTIC will run separate logistic regressions on each of the imputed datasets, since
we have used the “by _imputation_” option*/;

ods output parameterestimates=outlogistic_mcar25_lsay;

proc surveylogistic data=imputed;

model dropout = sex occ_asp escs math_std;

weight weight;

by _imputation_;

run;

ods output close;

*/ PROC MIANALYZE will pool results from the imputed datasets */;

*/ Also, we write the parameter estimates generated by proc mianalyze to a file called "ParmEst" */;

ods output mianalyze.parameterEstimates = ParmEst;

proc mianalyze parms(classvar=classval)=outlogistic_mcar25_lsay;

modeleffects intercept sex occ_asp escs math_std;

run;

ods output close;
```

VET COLLECTION

(no weights, uses PROC LOGISTIC only. Thus we use the “data=” option in PROC MI ANALYZE)

```
*/ PROC MI will create the specified number of imputed datasets */;

proc mi data = work.mcar25_lsay noprint nimpute=10 seed=12345 out=imputed;

var age disable no_yr12 cert_III_prior NESB dropout;

run;

*/ PROC LOGISTIC will run separate logistic regressions on each of the imputed datasets */;

proc logistic data=imputed outest=outlogistic_mcar25_lsay;

model   dropout (event='1') = age disable no_yr12 cert_III_prior NESB / rsquare;

by _imputation_;

run;

*/ PROC MIANALYSE will pool results from the imputed datasets */;

*/ Also, we write the parameter estimates generated by proc mianalyze to a file called "ParmEst" */;

ods output mianalyze.parameterEstimates = ParmEst;

proc mianalyze data=outlogistic_mcar25_vet;

modeleffects intercept age disable no_yr12 cert_III_prior NESB;

run;

ods output close;
```

NCVER

National Centre for Vocational Education Research Ltd
Level 11, 33 King William Street, Adelaide, South Australia
PO Box 8288, Station Arcade, SA 5000 Australia
Telephone +61 8 8230 8400   Facsimile +61 8 8212 3436
Website www.ncver.edu.au   Email ncver@ncver.edu.au