University of Pretoria
Department of Economics Working Paper Series

Individual and Household Debt: Does Imputation Choice Matter?

Adél Bosch
University of Pretoria and Development Bank of Southern Africa
Steven F. Koch
University of Pretoria
Working Paper: 2021-41
June 2021

Department of Economics
University of Pretoria
0002, Pretoria
South Africa
Tel: +27 12 420 2413

Individual and Household Debt

Does Imputation Choice Matter?

Adél Bosch[*] Steven F. Koch[†]

11 February, 2021

Abstract

In the case of sensitive requests - such as those made of survey respondents to reveal their earnings, their individual assets, debts or even net worth - complete answers are rarely forthcoming. Thus, there are numerous non-responses. We apply a bevy of imputation methods in an attempt to reduce the proportion of missing data on individual and household debt that is present in the National Income Dynamics Study panel data. Our application of Multiple Imputation by Chained Equation (MICE) yields additional observations on these variables than are available in the NIDS imputation data. Although our imputations do alter the distribution of the debt data across the first four waves, especially for individual level debt data, the effect of that alteration, once aggregated to the level of the household, is negligible.

1 Introduction

Missing responses in household survey data are unavoidable and present a challenge for research. Researchers can choose to (i) ignore missingness by dropping missing observations; (ii) recode missing information to certain values; (iii) recode, as in (ii), while including controls for missingness; and (iv) impute missing imputations via single and multiple imputation (MI) approaches. When data is missing completely at random (MCAR) and few observations are missing, the missing observations can be ignored with negligible impact on both bias and power (Graham 2009). If the missing share increases, but missing data remains MCAR, ignoring the missing observations reduces statistical power. However, MCAR is a strong assumption, and, therefore, ignoring missing data can lead to bias and power loss (Little and Rubin 1986). As such, it might be more beneficial to explore other options, such as imputing the missing data.

In what follows, we focus on MI methods to deal with missing wealth information in the National Income Dynamic Study (NIDS), which we compare to the regression imputations conducted by NIDS, as well as non-imputed data. Unlike single imputation methods, such as regression imputation, where the imputed

[*]Department of Economics, University of Pretoria and Development Bank of Southern Africa, AdelB@dbsa.org
[†]Corresponding author, Department of Economics, University of Pretoria, steve.koch@up.ac.za

value is treated as an actual known value, MI incorporates statistical uncertainty (Azur et al. 2011) allowing for several different imputed data sets to be created. Thus, MI can provide more robust information than what is initially available in the survey data or via a single imputation.[1]

Non-responses are typically high for sensitive items, such as: income, assets, and debt. In an attempt to gather information in the case of non-response, interviewers may provide alternative options, such as bracket responses. Bracket responses can provide a useful supplement to point value responses; they are solicited in an attempt to lower the non-response rate through queries that might be considered less sensitive by the respondent. In cases where respondents do not want to provide any values at all, interviewers may try further elicitation techniques. In the case of assets, wealth and debt in NIDS, interviewers try to determine if the household does have assets or debt. Binary responses (e.g., yes, we have debt), combined with additional survey data, are used to impute other missing values (e.g., the individual's or household's debt liability).

This research focuses primarily on income and debt, the former of which is likely to be an important determinant of the latter. Such data underpins any analysis of household debt portfolios, which are understudied (Zinman 2015). Similar data can also be used in analysis of deleveraging at the individual and/or household level, which is also an important component of comparative household finance (Zinman 2015; Badarinza, Campbell, and Ramadorai 2016). Furthermore, each of these variables, income and debt, incorporate many survey respondents only willing to share that they had earned income or held debt (or earned income or held debt in some range of values), but were not willing to provide a point value.

We use the first four waves of NIDS - a nationally representative panel survey - covering 2008, 2010/2011, 2012 and 2014/2015 (Southern Africa Labour and Development Research Unit 2016a, 2016b, 2016c, 2016d). Wave 2 and 4 contain comprehensive asset and debt questions. Since these questions ask about sensitive information many of the data are missing. All four wave household income questions allowed for bracket responses; however, for the wealth variables, most of the bracket options were only available in Wave 4.

We considered different imputation options, although our primary focus was on methods that have been applied in the South African literature, as well as methods that could take advantage of the binary response and bracket data that was available. Thus, we make use of multivariate imputation by chained equation (MICE), which is a useful method for dealing with missing data; MICE is also referred to as `fully conditional specification` or `sequential regression MI` in the literature. Additionally, MICE provides for bounds, which allows for the imputation of a point value from a bracket response. Given its common use in the South African literature (Posel and Casale 2005; Von Fintel 2007), we also considered mid-point methods.

[1] Ground-breaking work on non-response in household surveys was largely driven by Ferber (1966); DeMaio (1980); Lillard, Smith, and Welch (1986) and Little (1988).

Our analysis found little difference between mid-point and MI bracket methods.

However, much of the available South Africa data imputation research focuses on income (Posel and Casale 2005; Vermaak 2012; Von Fintel 2007; Wittenberg 2008) and income inequality (Wittenberg 2017a, 2017b). Although we do incorporate income, our main interest is debt, which has received less attention - Daniels, Finn, and Musundwa (2014) examine the wave 2 net wealth data, which therefore incorporates debt, relative to the South African Reserve Bank national accounts data and Daniels and Augustine (2016) provide a wave 4 update and similar comparison. We separate ourselves from previous South African debt research by (i) taking advantage of all of the waves, (ii) offering comparative insight across the waves, (iii) considering additional imputation methods and (iv) paying particular attention to both the individual and household debt data.

Through imputation, we were able to increase our point value observations for total debt by 1 467 in Wave 1, 823 in Wave 2, 1 051 in Wave 3 and 1 062 in Wave 4. We were also able to increase the household income response rate from 75% to 81% in Wave 1, 84% to 94% in Wave 2, from 90% to 97% in Wave 3 and from 95% to 98% in Wave 4. When imputing the individual debt variables for those who responded that held debt, we increased the total number of observations for all debt types by 43.8% in Wave 1, 35.4% in Wave 2 (including those that we derived from responses in Wave 1 and Wave 3, when the question was not asked in phase 2 of Wave 2), 26.3% in Wave 3 and 13.4% in Wave 4. We also show that imputing for non-responses did not change the averages of the debt variables or the distribution (except for the non-bank student loans in Wave 1 and Wave 2), which suggests that the data was initially MAR, and provides some support for the use of MICE.

2 Unfolding brackets and imputing item non-responses

Non-response rates are typically high for sensitive items such as income and wealth (Riphahn and Serfling 2005; Moore and Loomis 2001). Household members are either not comfortable providing the type of information required (Hurd 1999; Juster et al. 2007) or they do not know the exact information (Nwanzu 2010). Possibly, the respondent does not trust the interviewer (Riphahn and Serfling 2005). A characteristic feature of survey data on household wealth is the high incidence of missing data. Roughly, one in three respondents who report owning an asset are either unable or unwilling to provide an estimate of the value of the asset (Juster et al. 2007).

A partial solution to this problem is to devise a series of questions that put the respondent's value into a quantitative range. These quantitative ranges are called unfolding brackets, and they represent a survey innovation that substitutes range data for completely missing data (Juster et al. 2007). Although bracket values are not as good as point values, they provide a way to collect information from households not comfortable disclosing point values. Juster et al. (2007) and Hurd (1999) show that the information contained by a bracket response provides enough additional information to produce efficiency gains and reduce the imputation error.

The main contributors to unfolding brackets in the South African literature are Posel and Casale (2005) Vermaak (2012), Von Fintel (2007), Wittenberg (2008) and Wittenberg (2014). These authors use single and multiple imputation methods for determining point values from bracket responses via means, mid-points, conditional means and other methods. The mid-point method assigns the mid-point value of each bracket to the representative response in that bracket range (Wong et al. 2016). This procedure however introduces artificial spikes at the imputed values and is difficult to use when brackets are open-ended, especially at the top (where there is no obvious mid-point), although Von Fintel (2007) and Yu (2013) suggest that the top boundary can be quantified as 1.1 times the lower bracket boundary. In the published Stata code for income brackets, NIDS uses double the lower top-bracket for household income in Wave 2, 3 and 4, and 1.5 in Wave 1.

Apart from the mid-point method, Posel and Casale (2005) also test the actual average mean method, where all point values are divided into their corresponding brackets and the mean of the point values are then assigned to each of the brackets; see also (Casale 2004; Cichello, Leibbrandt, and Woolard 2012.) The authors find little difference between their average mean approach and the mid-point approach, except at higher and lower earnings brackets. Posel and Casale (2005) further use OLS regression and Heckman selection corrections, finding that it was important to account for bracket information. Von Fintel (2007) compares mid-point estimation to conditional mean imputation from both a lognormal and a pareto distribution, as well as interval regression, which incorporates earnings into a likelihood function. Similar to Posel and Casale (2005), Von Fintel (2007) shows that it is important to include bracket information in the approach. Therefore, we account for brackets in our analysis.

Evidence from Posel and Casale (2005) and Von Fintel (2007) suggest that mid-point values for earnings brackets perform similar to other more advanced methods. However, attention has been shifting away from deterministic imputation methods, like mid-points, towards methods taking cognisance of the imputation process and the underlying distribution of the data (Wittenberg 2014). When bracket information is not available, item non-responses can still be imputed to create a more complete set of analysis data (Daniell 2009). Methods include simple imputation, regression imputation, hot deck imputation, nearest neighbour imputation, predictive mean matching and MI methods (Durrant 2005). We make use of a combination of nearest neighbour, predictive mean matching and MI to impute missing values for the NIDS survey. We also compare our results to the publically provided NIDS data, where regression imputation is used.

Hot deck imputations replace missing values with observed values from a respondent that is similar with respect to selected characteristics (Andridge and Little 2010). Predictive mean matching (pmm) (Little 1988), on the other hand, imputes missing values by matching the nearest-neighbour, based on the expected values of the missing variables conditional on the observed characteristics (Vink et al. 2014). Specifically, observations are matched on the closeness of the predicted outcomes, but the imputed value will be the actual observed value of the matched neighbour (not the predicted value that was used for the match). The obvious implication is that imputation via pmm requires observed point values for matching, which may not always exist, for the survey of interest.[2]

Vermaak (2012) and Wittenberg (2014) use MI methods, which enables better estimates of the standard errors and more accurately reflects the underlying uncertainty in the imputation generating process. Vermaak (2012) determines the impact of MI on estimated poverty lines for South Africa. The method is used to impute both missing data and to impute missing data for restricted intervals, resulting in higher mean earnings, suggesting that the data were not missing at random. However, incorporating imputations did not result in a significant difference in the estimated poverty lines, when compared to ignoring missing values and using mid-points. Wittenberg (2014), however, finds that relying on one imputation could lead to bias and inconsistent estimates.

3 Methodology: multiple imputation and predictive mean matching (pmm)

MI was introduced by Little and Rubin (1986) and Rubin (1987), which transformed the discussion. Instead of replacing missing data or brackets with only one imputed estimate, MI allowed for several imputed values. The missing values are imputed based on the observed values for a given individual and the relations observed in the data for other individuals, assuming that the observed variables were included in the imputation model (Schafer and Graham 2002). Because MI involves creating multiple estimates for each missing value, the analysis of multiply imputed data must take into account uncertainty in the imputations to yield more accurate standard errors (Greenland and Finkle 1995).

According to Little and Rubin (1986), missingness can be defined in three ways:

[2] Wittenberg (2014) applied a similar approach, but used data from a different survey year, adjusted for inflation, rather than the survey being examined. Thus, the method is flexible.

1) Missing completely at random (MCAR) - The missing values do not depend on observed or unobserved variables. The missing cases can therefore be seen as randomly missing (Wayman 2003), and the only penalty is a loss of power due to fewer observations. This assumption is stronger than necessary and, in practise, it can be replaced with the more relaxed Missing at Random assumption.
2) Missing at Random (MAR) - The missing values depend on the observed data, can be fully described by other variables in the dataset and are not dependent on the missing data. This assumption underlies most imputation procedures (Nicoletti and Peracchi 2006). For example, even though respondents at the lower and upper end of the income distribution are less likely to provide survey responses than those in the middle, these missing data points are related to demographics and other socioeconomic variables, which can be observed in the data (Pedersen et al. 2017).
3) Missing not at random (NMAR) - The probability of missing values depend on unobserved data. This is also referred to as nonigorable missingness in the literature, while MCAR and MAR imply ignorable missingness.

According to Allison (2000), there are a few conditions that should be satisfied before MI can be used. These are:

a) The data must be missing at random (MAR).
b) We need to apply an estimator that matches the variable type; packages such as R's (R Core Team 2020) `mice` (Van Buuren and Groothuis-Oudshoorn 2011) automatically choose linear regression models for continues data and categorical response models for discrete data.
c) The model used for the analysis must match the model used in the imputation. It is suggested that the same variables used in the final regression should also be used in the imputation model (Allison 2002).

MICE can be implemented under the MAR assumption (Raghunathan et al. 2001; Van Buuren and Groothuis-Oudshoorn 2011), such that missing values depend only on observed information (Azur et al. 2011). Unfortunately there is no specific test for MAR (Kwak 2010); however, the `vim` package in R (Visualization and Imputation of Missing Values, Kowarik and Templ (2016)) is a useful tool to examine the correlation between missing data and other observed variables. MI generates multiple imputed values which replaces missing values, creating multiple unique data sets.

Suppose we have $X_1, X_2, ...X_p$ variables. If X_1 has missing values, we fit a model that is conditional on all the other variables, $X_2, ..., X_p$.[3] The missing value will then be replaced by the matched actual value

[3]A full description of the model can be found in Van Buuren et al. (2006); Van Buuren and Groothuis-Oudshoorn (2011) and Christelis (2011).

for pmm, or the predicted values for the other types of models that match the data. As implied above, we use automatic model selection for the imputation models. In other words, as each variable is imputed in turn, the model can be specified as either pmm for numeric variables, logistic regression for 2 factors, multinomial logit for 2+ factors or ordered logit for 2+ ordered factor. Similarly, if X_2 has missing values then $X_1, X_3...X_p$ variables will be used in the prediction model as independent variables.

In other words, for every variable in X_{-p} that precedes X_p in the sequence of variables, its value from iteration t is used (including the imputed values). Also, for every variable in X_{-p} that follows X_p in the sequence, its array of imputed values from iteration $t-1$ is used.

The steps can be summarised as follows:

1) Estimate a linear regression of the X's on X_p and produce a set of β coefficients using only observed values to estimate. $X_1 = X_2^t\beta_{12} + X_3^t\beta_{13} + ... + X_p^t\beta_{1p} + \varepsilon_1$ with $\varepsilon \sim \mathcal{N}(0, \sigma_1^2)$.
2) Draw a new set of coefficients β^* from its posterior distribution using predictive mean matching. Typically this is a random draw from $P(X)$, a multivariate normal distribution of X, with mean β and covariance matrix of β (with an additional draw for the residual variance); see Appendix A in Van Buuren et al. (2006) for more details.
3) Use β^* to generate predicted values for X for all cases (both for missing and non missing). 1)For each missing value of X match the predicted values of the observed values to the predicted values of the missing data.
4) Randomly select one of, in this case 5, nearest neighbour matches and assign the observed value to the missing data using the predicted matches.
5) Repeat step 2 to 5 for each complete data set. In our case, we repeat 10 times.

These steps are applied sequentially, and, after the imputation of the last variable, iteration t is considered complete. The iteration number can be between 10 and 20 Van Buuren and Groothuis-Oudshoorn (2011). Once the cycle is completed, multiple datasets are generated differing only in their imputed values.

Gibbs sampling, in which the parameters and missing values are drawn iteratively from appropriate conditional distributions, is used to obtain the joint posterior distribution of parameters and missing values given observed data. It is possible that the specification of two conditional distributions $P(X1|X2)$ and $P(X2|X1)$ are incompatible, so that no joint distribution $P(X1, X2)$ exists. Therefore, as the parameter vector of the joint distribution of X is replaced by the P different parameter vectors of the P conditional expectations, the posterior distribution is generated by a Gibbs sampler with data augmentation. To get convergence to the stationary distribution of X, we iterate the Gibbs sampler until we have a number of iterations indicating convergence of the distribution of the missing values of all variables in our system. Given that the imputations are sequential, we first impute the demographic variables, then the interaction variables and lastly our variable of interest. We set the visiting sequence in the MICE imputation specification.

4 Overview of the NIDS survey data and variables for imputation

NIDS is a South African panel that follows individuals across time (over different waves) and is considered to be representative of the South African population. NIDS is a comprehensive survey that covers topics regarding household finances such as income, debt, and assets. It further asks questions, among others, on poverty, well-being, labour markets, education, and health (NIDS 2018). The questionnaire covers households, adults, and children. We use the first four waves of the survey, conducted in 2008, 2010/2011, 2012, and 2014/2015. As our area of interest is household deleveraging, we focused on analysing data collected on household income and debt, as well as individual debt. In chapter 4 we attach these results to a household representative and follow them over time to establish to what extent households have deleveraged between wave 1 and the subsequent waves. Due to income bracket responses and the high incidents of missing values, for debt outstanding in the NIDS survey, we need to impute for these.

First, we impute for bracket responses provided in the household income question. The brackets differ across the waves and are imputed separately for each bracket and wave. Second, we impute the household level question on property debt outstanding. Third, we impute debt variables where individuals responded that they have debt, but do not provide a value. We impute for an anomaly in wave 2 regarding individual debt categories, where unlike in the other waves, NIDS rolled out two phases of interviews. In the first phase (phase 1) the questions regarding individual debt were asked. In the second phase (phase 2) these questions were not included in the interview (Daniels, Finn, and Musundwa 2014). We assign these phase 2 non-response values based on their wave 1 and wave 3 responses (i.e. if the respondent stated that they have debt in wave 1 and wave 3, they were assumed to have had debt in wave 2 and their outstanding debt value was imputed).

4.1 Unfolding household income brackets

NIDS report household income in point values, and in brackets. We unfold these brackets using MI and take the average of the 5 imputations as the point value for the bracket. The imputations are bounded by the bracket limits provided by NIDS. For the upper bound, where no upper limit is set in the survey, we use the 99.9th percentile of the point values as this allows the exclusion of some extreme values; see Table 1.

For the bracket imputations, we use *value of the house*, *mode of the household ethnicity*, *maximum level of education in the household*, *median age of the household*, *province* and *household size*. The selection of household demographic transformations (mode, maximum and median) follows the approach used by NIDS for household imputations. However, in this case, NIDS used mid-points for the income bracket imputations and unfolded the upper bound by doubling it in Wave 2, 3 and 4 and multiplying it by 1.5 in Wave 1. We show that using either of the two methods does not make a large difference in the result of of the income variable; see Figure 1.

The results show that we increase the household income response rate from 75% to 81% in Wave 1, 84% to 94% in Wave 2, from 90% to 97% in Wave 3 and from 95% to 98% in Wave 4; see Table 2. Similar to results found in the South African literature for household survey income brackets (Posel and Casale 2005; Von Fintel 2007; Wittenberg 2008, 2014; Vermaak 2012), we also confirm that using mid-points or MI produces similar distributions for NIDS data. The mean household income from using MI is consistently lower than those for mid-points, suggesting that the mid-point method perhaps overstates incomes in the bracket; see Table 3.

To test if the Gibbs sampler converged, we follow Van Buuren and Groothuis-Oudshoorn (2011) and plot the mean and standard deviations against their iterations. Figure 2 plots the Gibbs sampler for each of the upper bounds of the income brackets. When plotting the Gibbs sampler we can determine convergence by inspecting the different sequences. On convergence, the sequences should be freely intermingled with each other, without showing any definite trends. Convergence occurs when the variance between different sequences is not larger than the variance within each individual sequence (Van Buuren and Groothuis-Oudshoorn 2011). We also confirm convergence using the (Gelman and Rubin 1992) test. This test provides is based on the *Rhat* statistic – approximately the square root of the variance of the mixture of all chains, divided by the average variance within the chain (Gelman and Hill 2007). We used the more stringent criterion (Rhat ≤ 1.1), rather than a less stringent criteria of Rhat ≤ 1.2 (Su et al. 2011). Intuitively, if Rhat is much greater than 1, the chains have not mixed well. In our case, apart from the graphical confirmation, Rhat did not exceed 1.1 for any income bracket parameters, while the upper brackets for household income also showed healthy Gibbs sampler convergence, Figure 3.

4.2 Imputing debt outstanding

This section evaluates the imputations of debt outstanding for the households and individuals that reported that they did have a certain debt type, but provided no value for the amount outstanding (they gave a yes/no answer to the question on whether they had debt outstanding). NIDS also provide code for regression imputations. However, their regression imputations have limitations (e.g. imputations are only done for variables if there are at least 100 observations and more than a 40% combined observation and participation rate). For MI, we do not have these limitations. However, unlike regression imputation, we can have multicollinearity when including too many similar demographic variables in the prediction matrix, as they are all used at the same time and may be linearly related and the matrix will be computationally singular and not solve. However, by setting the diagonals of the predictor matrix to zero it eliminates the possibility of not solving, as duplicate information is excluded when imputing. Furthermore, for MI, if there are too few observations, such as only one, imputations will fail, however we did not have any examples of this in our data.

Table 4 is a summary of the variables imputed and the "X" indicate for which waves of the survey the variables were available. The first variable we impute is a household level variable. It is based on a set of survey questions attempting to determine if the respondent's property is fully paid off, and if not, and the amount of outstanding debt was not provided, it was imputed. Similarly, based on the survey questions, we imputed for individual who responded that they had debt outstanding for the different individual debt categories as shown in Table 4, but did not give an amount outstanding. As shown in the Table 4 a separate question for if the respondent has a loan from a friend or a family member was only asked in Wave 2 and Wave 4, while the loan category was combined in Wave 1 and Wave 3. These were imputed in the same way.

4.3 Household property debt outstanding

For property debt outstanding NIDS has an individual and household level question. We impute both the household level property debt outstanding as well as the individual property debt outstanding. Individual property debt outstanding is discussed with the other individual debt outstanding variables in the next section. We impute for households who said that they owned the dwelling but it was not fully paid off, however they did not provide an amount outstanding. See Figure 4 for how the questions were asked in the survey.

By imputing for households who reported owing money on a dwelling, but did not provide a value, the number of responses nearly doubles; see Table 5. Average property debt is also reduced, Table 6. In Wave 4, regression imputation results in even lower average property debt than the MI impute procedure. The differences in the kernel densities between the MI, regression imputations and no imputations are illustrated in Figure 5, while the Gibbs sample convergence is available in Figure 6 for household debt. Because the sequences freely intermingle, we confirm a healthy convergence. The Gelman and Rubin (1992) test result also confirms convergence; all Rhat statistics are below 1.1, Figure 7.

4.4 Individual debt outstanding

As mentioned earlier, there was the two-phase anomaly in Wave 2, where NIDS conducted a phase 1 interview, asking if individuals have debt, while not asking the respondents in phase 2 these questions (Daniels, Finn, and Musundwa 2014), resulting in a lower response rate. In order to increase Wave 2's response rate, we use information from Wave 1 and Wave 3 to infer whether the person also has this type of debt in Wave 2. These were then imputed.[4] Also, see Daniels and Augustine (2016) for an overview of the NIDS wealth, including debt variables for Wave 4. In the next section, we describe the imputations for individual debt questions. Figures 8 show how the questions are asked in the questionnaire. Ownership of a vehicle is used as an example.

The responses for Wave 1, Wave 3 and Wave 4 are shown in Table 7, while the responses, with a focus on the two-phase interview, are shown in Table 8. For Wave 1, 2 and 4, we impute a debt outstanding value from those who responded yes to the question on do they have debt, but provided no debt amount outstanding. For Wave 2, we firstly inferred that those who were not asked the question in phase 2 had debt based on their debt status in Wave 1 and Wave 3. We then imputed the amount of debt outstanding based on the answers from phase 1 and the inferred answers from phase 2.

We report the imputations for individual total outstanding debt on the dwelling owned. We use individual level information and the household income variable for imputations. Individual information consists of *ethnic group*, *gender*, *education*, *age*, *age squared*, *ethnic/gender interactions*, *married* and *province*. We use the mean of the unfolded brackets for income and impute the additional debt participation variables, where income is an independent variable. We allow for credit, store card and bond debts to be zero.

Figures @ref(fig:fig2.5) to @ref(fig:fig2.8) show the different distributions for the debt variables for each wave of the survey. For Wave 1 and Wave 3 there were no NIDS imputations, so the results only show the MI imputations versus no imputations, while for Wave 4, where NIDS imputations are available, the results are compared to no imputations and MI imputations. The distributions look fairly similar, however, for Wave 1 and Wave 2, the imputations for student loans from a non-bank institution for the MI are more negatively skewed. The Gelman and Rubin (1992) test results are shown in Figure 13 and shows convergence of the Gibbs sampler for all individual debt imputations, while the Gibbs sampler convergence graphs are available from the author.

When adding all the actual point responses for the individual debt variables across the waves, see Table 9,

[4] It is possible that for very short-term debt, households may have deleveraged fully between the waves. However, we only make this assumption for a combined 31 individuals with store and credit card debt, and, therefore, we do not believe the assumption is consequential with respect to total imputations.

we find that by imputing values, as described above, we lifted the number of observations, in some instances even by double. For Wave 1 we increase the number of responses from 3 350 to 4 817. For Wave 2, it is from 2 323 to 3 146, which is more than the NIDS imputations that lifts the number only to 2 984 because, as mentioned earlier, we impute for the debt participation questions that were not asked in the second phase of Wave 2. For Wave 3, we manage to improve the response number from 4 003 to 5 054, and for Wave 4 from 7 952 to 9 014 when using MI.

We also show that imputing for non-responses does not change the averages (Table 10 of the debt variables or their distributions (except for the non-bank student loans in Wave 1 and Wave 2), which implies that the data was initially MAR and, therefore, imputations provide additional data points. Thus, by using MI, information regarding the distribution of the imputations is left intact when utilising regression analysis.

5 Conclusion

We used the first four waves of NIDS, which are from the years: 2008, 2010/2011, 2012 and 2014/2015. The surveys were used to determine individual and household level debt. However, since such information is sensitive, we used MICE to impute missing data and point values from bracket responses. The primary advantage of MICE is that it allows for bounds imputation; in other words, imputed values are forced to lie within a required range, and such ranges are available from bracket responses. Such data can be useful to examine household debt portfolios, which are understudied (Zinman 2015), as well as deleveraging at the individual and household level, which is also an important component of comparative household finance (Zinman 2015; Badarinza, Campbell, and Ramadorai 2016). We focused our imputations on income and household debt - the former of which is likely to be an important determinant of the latter. Furthermore, each of these variable types include many who responded that they had earned income or held debt, but did not provide a point value. Thus, we used the information to impute a point value, where possible, and we compared that data to imputations available directly from NIDS, as well as the non-imputed data.

We established that, as is typical of household surveys and already established with respect to NIDS (Daniels, Finn, and Musundwa 2014; Daniels and Augustine 2016), that the incidence of non-response to sensitive questions, such as on income, assets and debt, were quite high. Upon imputation, we were able to capture a few more individuals and households in the underlying debt and income data than were initially captured through NIDS. The additional data did affect the underlying distributions, and, therefore, the associated means and densities.

Given the need for imputation, we considered different options available, although our primary focus was on methods that have been applied in the South African literature, as well as methods that could take advantage of the binary response and bracket data that was available. Thus, we incorporated both mid-points and MI methods, for dealing with bracket data. Our analysis found little difference between mid-point unfolding and MI methods. However, we were able to increase the household income response rate from 75% to 81% in Wave 1, 84% to 94% in Wave 2, from 90% to 97% in Wave 3 and from 95% to 98% in Wave 4. When imputing the individual debt variables for those who responded that they were debt participants, we increased the total number of observations for all debt types by 43.8% in Wave 1, 35.4% in Wave 2 (including those that we derived from responses in Wave 1 and Wave 3, when the question was not asked in phase 2 of Wave 2), 26.3% in Wave 3 and 13.4% in Wave 4. We also show that imputing for non-responses did not change the averages of the debt variables or the distribution (except for the non-bank student loans in Wave 1 and Wave 2), which suggests that the data was initially MAR, and provides some support for the use of MICE.

References

Allison, Paul D. 2000. "Multiple Imputation for Missing Data: A Cautionary Tale." *Sociological Methods and Research* 28: 301–9. https://www.sas.upenn.edu/%7B~%7Dallison/MultInt99.pdf.

———. 2002. "Missing Data." *Quantitative Applications in the Social Sciences* 136: 104. https://doi.org/10.1136/bmj.38977.682025.2C.

Andridge, Rebecca R., and Roderick J A Little. 2010. "A review of hot deck imputation for survey non-response." *International Statistical Review* 78 (1): 40–64. https://doi.org/10.1111/j.1751-5823.2010.00103.x.

Azur, Melissa J., Elizabeth A. Stuart, Constantine Frangakis, and Philip J. Leaf. 2011. "Multiple Imputation by Chained Equations: What Is It and How Does It Work?" *International Journal of Methods in Psychiatric Research* 20 (1): 40–49. https://doi.org/10.1002/mpr.329.

Badarinza, Cristian, John Y. Campbell, and Tarun Ramadorai. 2016. "International Comparative Household Finance." *Annual Review of Economics* 8: 111–44. https://doi.org/https://doi.org/10.1146/annurev-economics-080614-115640.

Casale, Daniela. 2004. "What has the feminisation of the labour market bought women in South Africa? Trends in labour force participation, employment and earnings, 1995–2001." *Journal of Interdisciplinary Economics* 15 (3-4): 251–75. https://doi.org/10.1177/02601079X04001500302.

Christelis, Dimitrios. 2011. "Imputation of Missing Data in Waves 1 and 2 of SHARE." CSEF Working Papers 278. Centre for Studies in Economics; Finance (CSEF), University of Naples, Italy. https://ideas.repec.org/p/sef/csefwp/278.html.

Cichello, P., M. Leibbrandt, and I. Woolard. 2012. *Southern Africa Labour and Development Research Unit Education: Analysis of the NIDS Wave 1 and 2 Datasets*. 78. http://www.nids.uct.ac.za/home/index.php?/Nids-Documentation/discussion-papers.html.

Daniell, Henry. 2009. "The use of sample weights in hot deck imputation." *National Instittute of Health* 76 (October 2009): 211–20. https://doi.org/10.1007/s11103-011-9767-z.Plastid.

Daniels, Reza C, and Taryn Augustine. 2016. "The measurement and distribution of household wealth in South Africa using the National Income Dynamics Study (NIDS) Wave 4." Working Paper 183. SALDRU Working Paper. Southern Africa Labour; Development Research Unit. www.saldru.uct.ac.za.

Daniels, Reza C, Arden Finn, and Sibongile Musundwa. 2014. "Wealth Data Quality in the National Income Dynamics Study Wave 2." *Development Southern Africa* 31 (1): 31–50. https://doi.org/10.1080/0376835X.2013.858308.

DeMaio, Theresa J. 1980. "Refusals: Who, Where and Why." *Public Opinion Quarterly* 44 (2): 223–33. https://doi.org/10.1086/268586.

Durrant, Gabriele B. 2005. "Imputation Methods for Handling Item- Nonresponse in the Social Sciences. ESRC National Centre for Research Methods and Southampton Statistical Sciences Research Institute." *NCRM Methods Review Papers* NCRM/002.

Ferber, Robert. 1966. "Item Nonresponse in a Consumer Survey." *Public Opinion Quarterly* 30 (3): 399–415. https://doi.org/10.1086/267432.

Gelman, Andrew, and Jennifer Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://books.google.co.za/books/about/Data%7B/_%7DAnalysis%7B/_%7DUsing%7B/_%7DRegression%7B/_%7Dand%7B/_%7DMulti.html?id=lV3DIdV0F9AC%7B/&%7Dredir%7B/_%7Desc=y.

Gelman, Andrew, and Donald B. Rubin. 1992. "Inference from Iterative Simulation Using Multiple Sequences." *Statistical Science* 7 (4): 457–72. https://doi.org/10.1214/ss/1177011136.

Graham, John W. 2009. "Missing Data Analysis: Making It Work in the Real World." *Annual Review of Psychology* 60 (1): 549–76. https://doi.org/10.1146/annurev.psych.58.110405.085530.

Greenland, Sander, and William D. Finkle. 1995. "A Critical Look at Methods for Handling Missing Covariates in Epidemiologic Regression Analyses." *American Journal of Epidemiology* 142 (12): 1255–64. https://doi.org/10.1093/oxfordjournals.aje.a117592.

Hurd, Michael D. 1999. "Anchoring and Acquiescence Bias in Measuring Assets in Household Surveys." *Journal of Risk and Uncertainty* 19 (1-3): 111–36. https://ideas.repec.org/a/kap/jrisku/v19y1999i1-3p111-36.html.

Juster, F., Honggao Cao, Mick Couper, Daniel Hill, Michael Hurd, Joseph Lupton, Michael Perry, and James Smith. 2007. "Enhancing the Quality of Data on the Measurement of Income and Wealth." Working Papers. University of Michigan, Michigan Retirement Research Center. https://EconPapers.repec.org/RePEc:mrr:papers:wp151.

Kowarik, Alexander, and Matthias Templ. 2016. "Imputation with the R Package VIM." *Journal of Statistical Software* 74 (7): 1–16. https://doi.org/10.18637/jss.v074.i07.

Kwak, Do Won. 2010. "Tests of MCAR and estimation methods for missing data: An application to the effect of class size reduction on SAT score for grades in K-3. Michigan State University." Job market paper. http://www.uq.edu.au/economics/documents/jobmarketpapers/jmpkwak.pdf.

Lillard, Lee, James P Smith, and Finis Welch. 1986. "What Do We Really Know about Wages? The Importance of Nonreporting and Census Imputation." *Journal of Political Economy* 94 (3, Part 1): 489–506.

Little, Roderick J A. 1988. "Missing-Data Adjustments in Large Surveys." *Journal of Business and Economic Statistics* 6 (3): 287–96. https://doi.org/10.2307/1391878.

Little, Roderick J A, and Donald B Rubin. 1986. *Statistical Analysis with Missing Data*. New York, NY, USA: John Wiley & Sons, Inc.

Moore, Jeffrey C, and Laura S Loomis. 2001. "Using alternative question strategies to reduce income nonresponse. Statistical Research Division U.S. Bureau of the Census Washington D.C. 20233." Research Report Series. https://pdfs.semanticscholar.org/5c30/affe397210374521bd23b147e67be63451ea.pdf%20https://www.census.gov/srd/papers/pdf/rsm2001-03.pdf.

Nicoletti, Cheti, and Franco Peracchi. 2006. "The Effects of Income Imputation on Microanalyses: Evidence from the European Community Household Panel." *Journal of the Royal Statistical Society: Series A*

(Statistics in Society) 169 (3): 625–46. https://doi.org/10.1111/j.1467-985X.2006.00421.x.

NIDS. 2018. "Frequently asked questions about NIDS." [Available:] ttp://www.nids.uct.ac.za/nids-data/documentation/faqs/data-about-nids.

Nwanzu, Godwin. 2010. "Factors , Preventions and Correction Methods for Non-Response in Sample Surveys." *Journal of Applied Statistics* 1 (1): 79–89. https://www.cbn.gov.ng/OUT/2012/PUBLICATIONS/REPORTS/STD/FACTORS,%20PREVENTIONS%20AND%20CORRECTION%20METHODS%20FOR%20NON-RESPONSE%20IN%20SAMPLE%20SURVEYS.PDF.

Pedersen, Alma, Ellen Mikkelsen, Deirdre Cronin-Fenton, Nickolaj Kristensen, Tra My Pham, Lars Pedersen, and Irene Petersen. 2017. "Missing data and multiple imputation in clinical epidemiological research." *Clinical Epidemiology* Volume 9 (March): 157–66. https://doi.org/10.2147/CLEP.S129785.

Posel, D, and D Casale. 2005. "Who replies in brackets and what are the implications for earnings estimates? An analysis of earnings data from South Africa." Working Papers 07. Economic Research Southern Africa. https://ideas.repec.org/p/rza/wpaper/07.html.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Raghunathan, T E, J M Lepkowski, J Van Hoewyk, and P Solenberger. 2001. "A multivariate technique for multiply imputing missing values using a sequence of regression models." *Survey Methodology* 27 (1): 85–95.

Riphahn, Regina T., and Oliver Serfling. 2005. "Item non-response on income and wealth questions." *Empirical Economics* 30 (2): 521–38. https://doi.org/10.1007/s00181-005-0247-7.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, Joseph L, and John W Graham. 2002. "Missing Data: Our View of the State of the Art." *Psychological Methods* 7 2: 147–77.

Southern Africa Labour and Development Research Unit. 2016a. "National Income Dynamics Study 2008, Wave 1." Cape Town: DataFirst [distributor], 2016: [Available:] http://www.nids.uct.ac.za/nids-data/data-access; Cape Town: Southern Africa Labour; Development Research Unit [producer], 2016.

———. 2016b. "National Income Dynamics Study 2010-2011, Wave 2." Cape Town: DataFirst [distributor], 2016: [Available:] http://www.nids.uct.ac.za/nids-data/data-access; Cape Town: Southern Africa Labour; Development Research Unit [producer],2012.

———. 2016c. "National Income Dynamics Study 2012, Wave 3." Cape Town: DataFirst [distributor], 2016: [Available:] http://www.nids.uct.ac.za/nids-data/data-access; Cape Town: Southern Africa Labour; Development Research Unit [producer], 2016.

———. 2016d. "National Income Dynamics Study 2014 - 2015, Wave 4." Pretoria: Department of Planning Monitoring; Evaluation [commissioner], 2014: [Available:] http://www.nids.uct.ac.za/nids-data/data-access; Cape Town: Southern Africa Labour; Development Research Unit [producer], 2016.

Su, Yu-Sung, Andrew Gelman, Jennifer Hill, and Masanao Yajima. 2011. "Multiple Imputation with Diagnostics in R : Opening Windows into the Black Box." *Journal of Statistical Software* 45 (2): 1–31. https://doi.org/10.18637/jss.v045.i02.

Van Buuren, Stef, J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D B Rubin. 2006. "Fully conditional specification in multivariate imputation." *Journal of Statistical Computation and Simulation* 76 (12): 1049–64. https://doi.org/10.1080/10629360600810434.

Van Buuren, Stef, and Karin Groothuis-Oudshoorn. 2011. "Mice: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software* 45 (3): 1–67. https://www.jstatsoft.org/v45/i03/.

Vermaak, Claire. 2012. "Tracking poverty with coarse data: evidence from South Africa." *The Journal of Economic Inequality* 10 (2): 239–65. https://doi.org/10.1007/s10888-011-9211-2.

Vink, Gerko, Laurence E. Frank, Jeroen Pannekoek, and Stef Van Buuren. 2014. "Predictive mean matching imputation of semicontinuous variables." *Statistica Neerlandica* 68 (1): 61–90. https://doi.org/10.1111/stan.12023.

Von Fintel, Dieter. 2007. "Dealing with earnings bracket responses in household surveys? How sharp are midpoint imputations?" *The South African Journal of Economics* 75 (2): 293–312. https://doi.org/10.1111/j.1813-6982.2007.00122.x.

Wayman, Jeffrey C. 2003. "Multiple Imputation For Missing Data: What Is It And How Can I Use It?" Paper presented at the annual meeting of the American Educational Research Association. Chicago. www.csos.jhu.edu.

Wittenberg, Martin. 2008. "Nonparametric estimation when income is reported in bands and at points." Working Paper 94. Economic Research Southern Africa. https://ideas.repec.org/p/rza/wpaper/94.html.

———. 2014. "Wages and wage inequality in South Africa 1994-2011: The evidence from household survey data." SALDRU Working Paper 135. Southern Africa Labour; Development Research Unit, University of Cape Town. https://ideas.repec.org/p/ldr/wpaper/135.html.

———. 2017a. "Wages and Wage Inequality in South Africa 1994–2011: Part 1 – Wage Measurement and Trends." *South African Journal of Economics* 85 (2): 279–97. https://doi.org/https://doi.org/10.1111/saje.12148.

———. 2017b. "Wages and Wage Inequality in South Africa 1994–2011: Part 2 – Inequality Measurement and Trends." *South African Journal of Economics* 85 (2): 298–318. https://doi.org/https://doi.org/10.1111/saje.12147.

Wong, Rebeca, Karina Orozco-Rocha, Dong Zhang, Alejandra Michaels-Obregon, and Cesar Gonzalez-Gonzalez. 2016. "Imputation of Non-Response on Economic Variables in the Mexican Health and Aging Study (MHAS/ENASEM) 2012." Project Report. www.MHASweb.org.

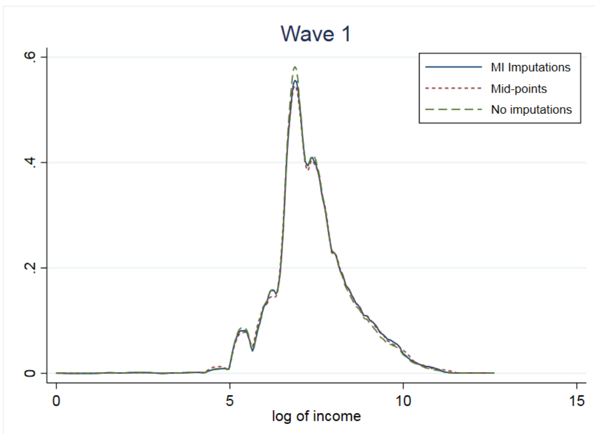
Yu, Derek. 2013. "Some factors influencing the comparability and reliability of poverty estimates across household surveys." *Stellenbosch Economic Working Papers* 03/13.

Zinman, Jonathon. 2015. "Household Debt: Facts, Puzzles, Theories, and Policies." *Annual Review of Economics* 7: 251–76. https://doi.org/https://doi.org/10.1146/annurev-economics-080614-115640.

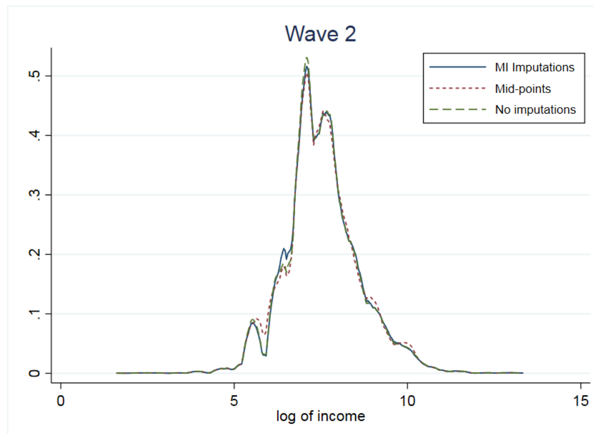
Table 1: Bracket responses for household income by wave

Wave 1	Wave 2	Wave 3	Wave 4
R 1 - R 200	R 0 - R 699	R 0 - R 499	R 0-R 749
R 201 - R 500	R 701 - R 1 299	R 501 - R 1 399	R 751 - R 1 499
R 501 - R 1 000	R 1 301 - R 2 299	R 1 401 - R 2 999	R 1 501 - R 2 999
R 1 001 - R 1 500	R 2 301 - R 4 699	R 3 001 - R 4 499	R 3 001 - R 5 999
R 1 501 - R 2 500	R 4 701 - R 10 999	R 4 501 - R 9 499	R 6 001 - R 10 999
R 2 501 - R 3 500	R 11 000 - R 229 121.6	R 9 501 - R 23 499	R 11 001 - R 26 999
R 3 501 - R 4 500		R 23 500 - R 236 850	R 27 001 - R 106 055
R 4 501 - R 6 000			
R 6 001 - R 8 000			
R 8 001 - R 11 000			
R 11 001 - R 16 000			
R 16 001 - R 30 000			
R 30 001 - R 50 000			
R 50 001 - R 88 975			

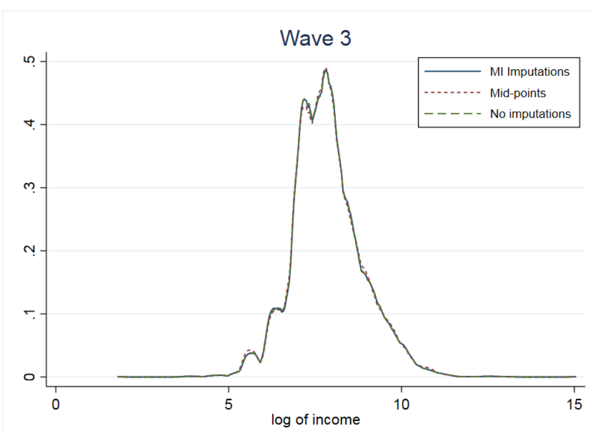
Upper brackets are the 99.9th percentile of the point values. For Wave 2 to Wave 4 the in-between values were provided as point values in the question and therefore used as point values and not imputed.



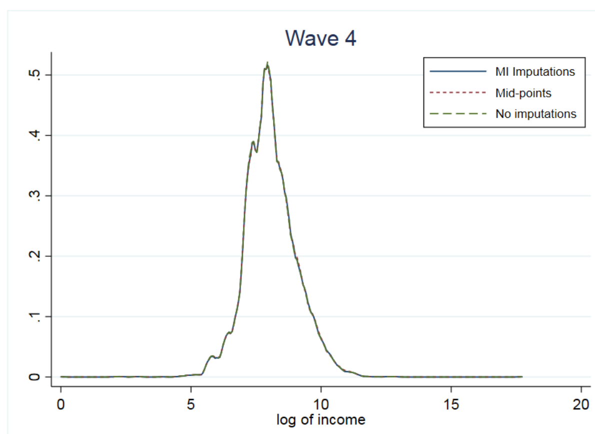
(a) Wave 1



(b) Wave 2

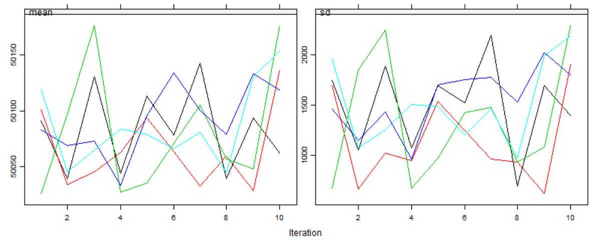


(c) Wave 3

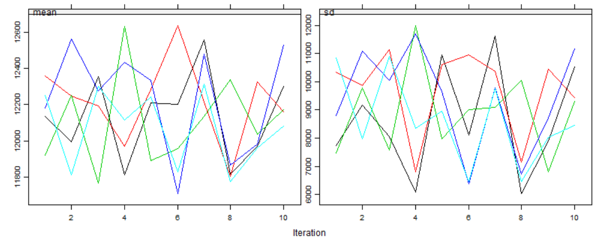


(d) Wave 4

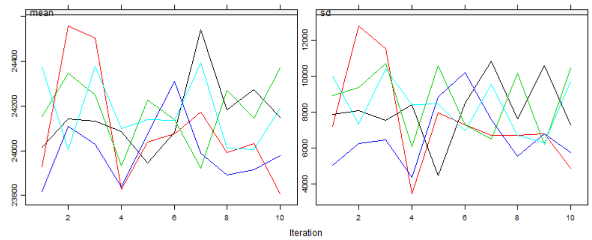
Figure 1: Kernel densities comparing multiple imputation and mid-points for unfolding household income brackets



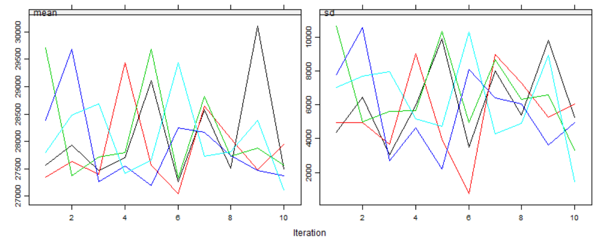
(a) Wave 1: Household income



(b) Wave 2: Household income

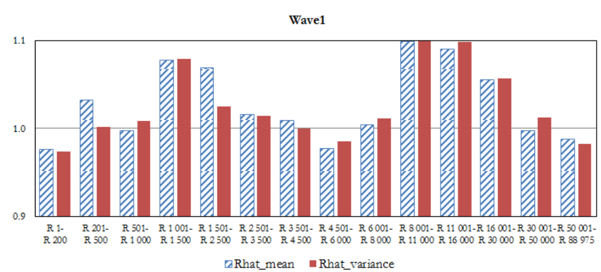


(c) Wave 3: Household income

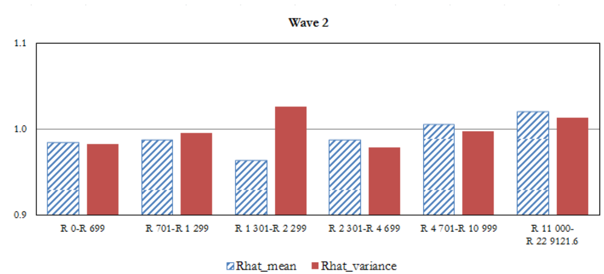


(d) Wave 4: Household income

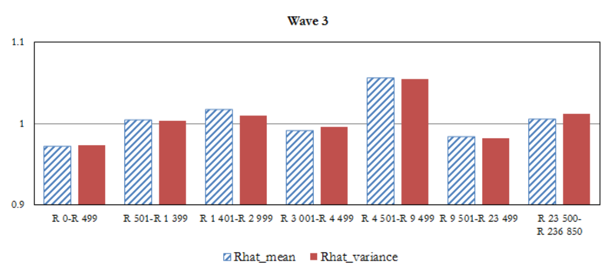
Figure 2: Convergence of Gibbs sampler for the upper bracket of household income



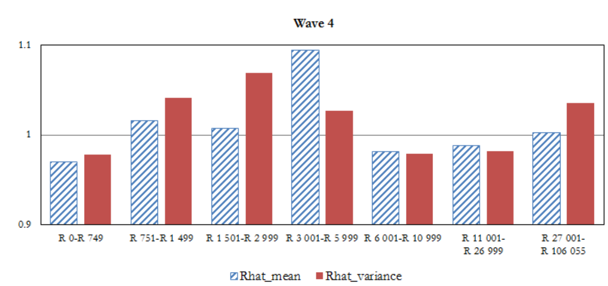
(a) Wave 1



(b) Wave 2



(c) Wave 3



(d) Wave 4

Figure 3: Rhat for convergence of Gibbs sampler for the upper bracket of household income

D4	Does a household member own this dwelling?	
	Yes	1
	No → **SKIP TO D10**	2

(a) Property ownership

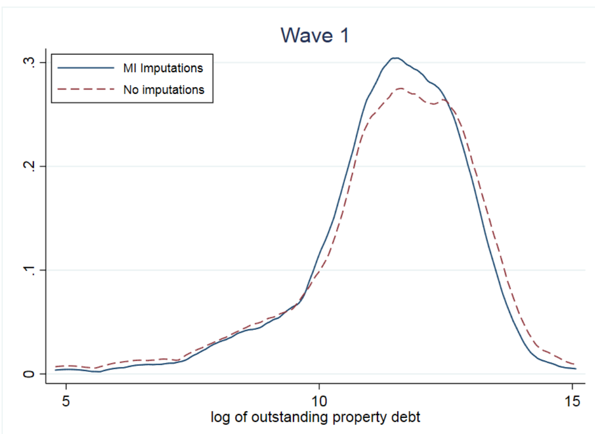
D6	Is this property fully paid off?	
	Yes → **SKIP TO D9**	1
	No	2
	Refuse	-8

(b) Property paid

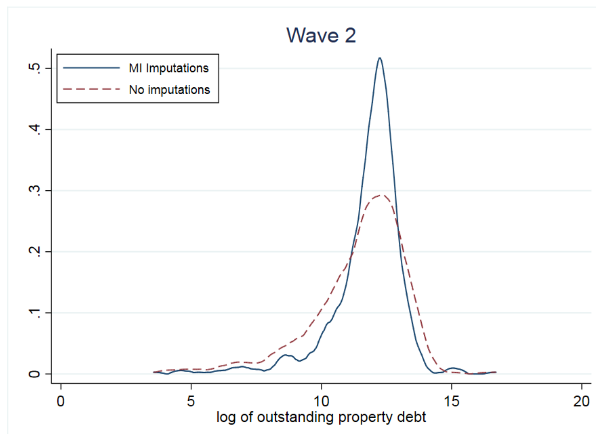
D7	What is the amount of the bond still owing on this property?	
	Amount	**R**
	Refuse	-8
	Don't know	-9

(c) Outstanding mortgage debt

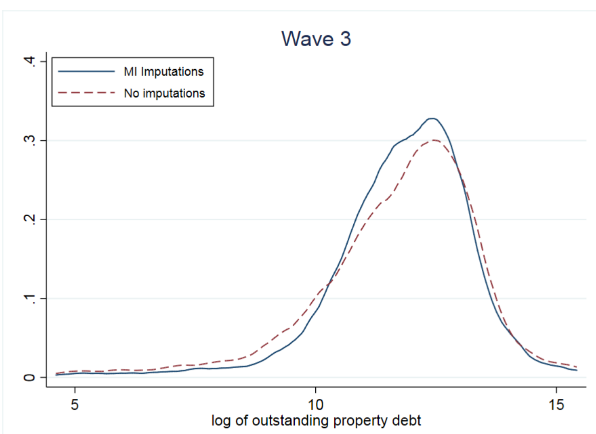
Figure 4: NIDS question on household debt outstanding on property



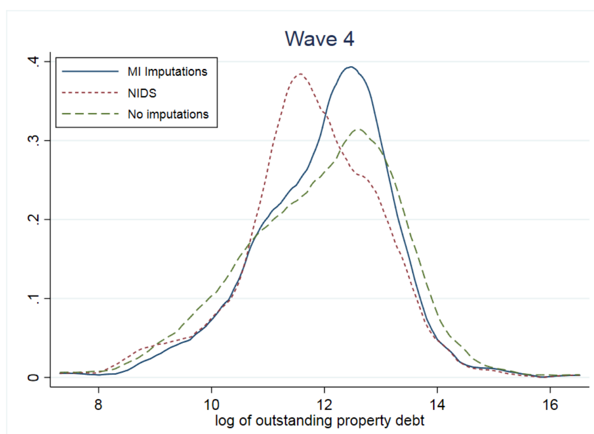
(a) Wave 1



(b) Wave 2

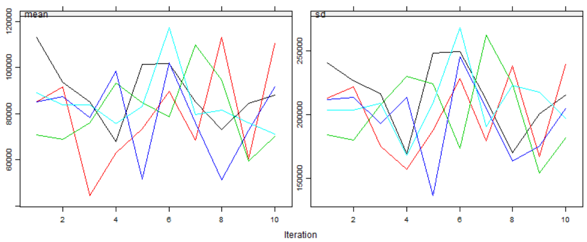


(c) Wave 3

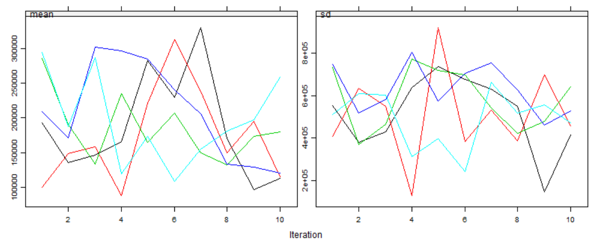


(d) Wave 4

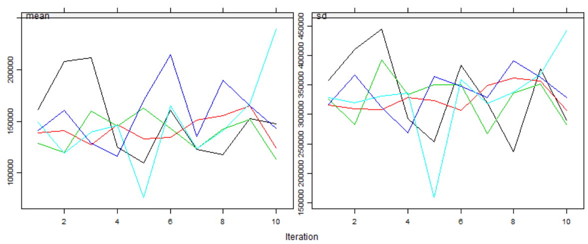
Figure 5: Kernel densities comparing multiple imputation and NIDS imputed (where applicable) for other household property debt outstanding



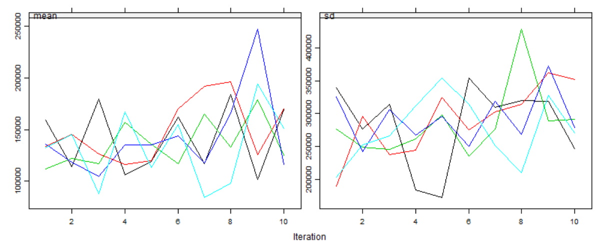
(a) Wave 1



(b) Wave 2



(c) Wave 3



(d) Wave 4

Figure 6: Convergence of Gibbs sampler for household debt

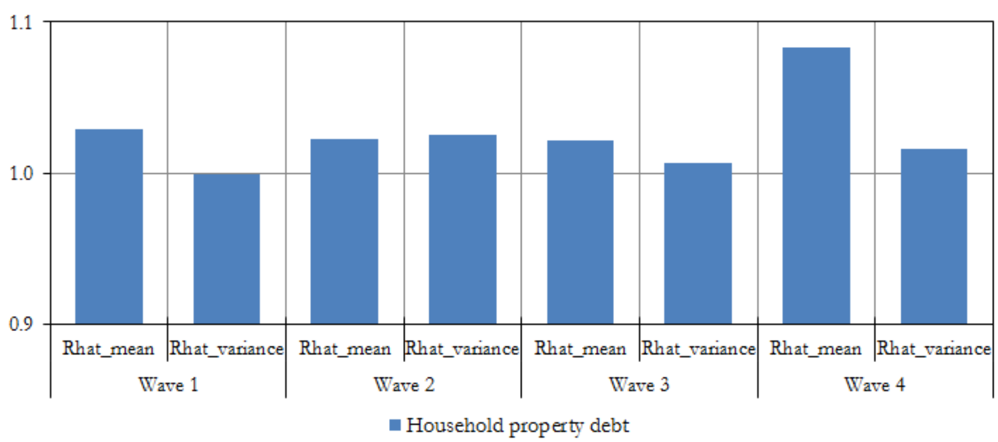
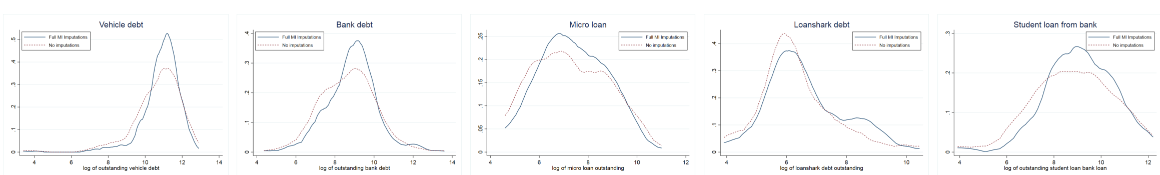


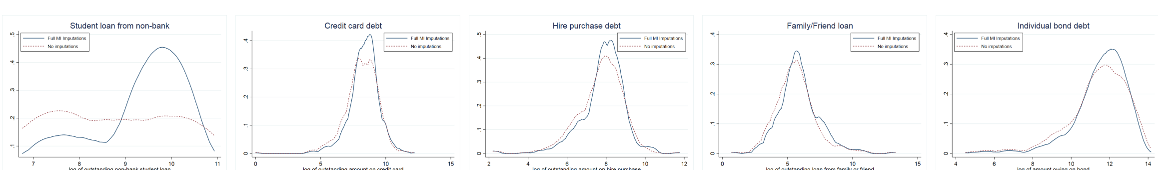
Figure 7: Rhat for convergence of Gibbs sampler for household debt

		1 Do you personally have a [...]? **Interviewer: If no, → SKIP TO NEXT** Yes	 No	2 What was the value of your payment on your [...] last month? **Interviewer: If don't know, write -9 If none, write 0** Rands	3 What is the remaining outstanding balance on your [...]? **Interviewer: If don't know, write -9 If none, write 0** Rands
G11	Home loan / Bond	1	2		
G12	Personal loan from a bank	1	2		
G13	Personal loan from a micro-lender	1	2		
G14	Loan with a Mashonisa	1	2		
G15	Study loan with a bank	1	2		
G16	Study loan with an institution other than a bank	1	2		
G17	Vehicle finance (car payment)	1	2		
G18	Credit card	1	2		
G19	Store card (For example, Edgars, Foschini or Woolworths store card)	1	2		
G20	Hire purchase agreement	1	2		
G21	Loan from a family member or friend	1	2		

Figure 8: Rhat for convergence of Gibbs sampler for household debt

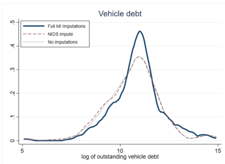


(a) Vehicle (b) Bank (c) Micro Loans (d) Masho? (e) Study

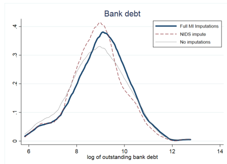


(f) Study Other (g) Credit (h) Hire (i) Family/Friend (j) Bond

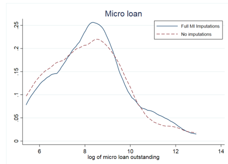
Figure 9: Wave 1 imputations



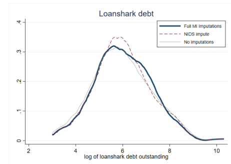
(a) Vehicle



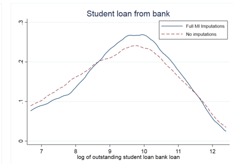
(b) Bank



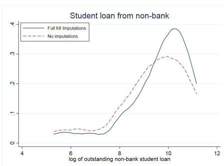
(c) Micro Loans



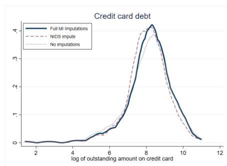
(d) Masho?



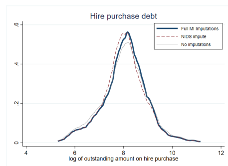
(e) Study



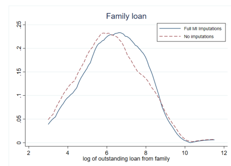
(f) Study Other



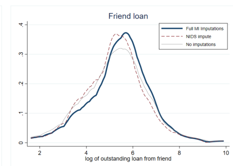
(g) Credit



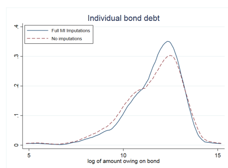
(h) Hire



(i) Family

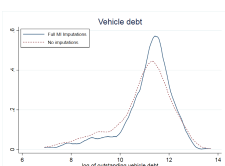


(j) Friend

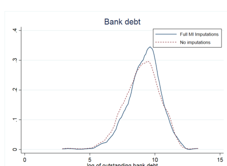


(k) Bond

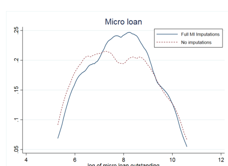
Figure 10: Wave 2 imputations compared to NIDS imputations (when applicable)



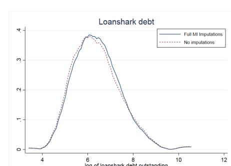
(a) Vehicle



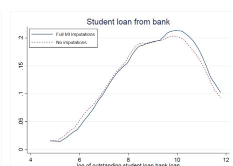
(b) Bank



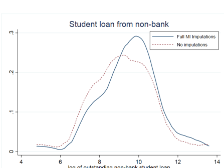
(c) Micro Loans



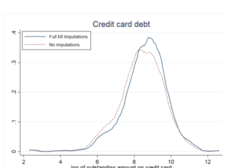
(d) Masho?



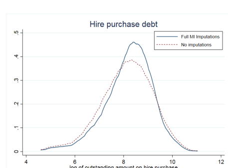
(e) Study



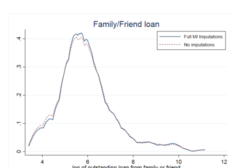
(f) Study Other



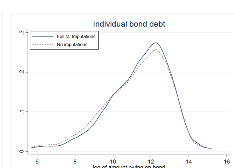
(g) Credit



(h) Hire

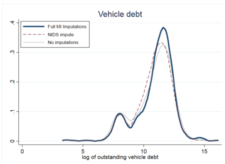


(i) Family/Friend

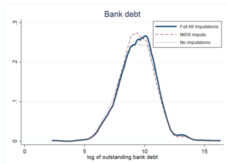


(j) Bond

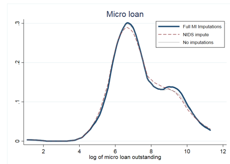
Figure 11: Wave 3 imputations



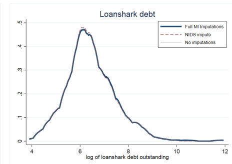
(a) Vehicle



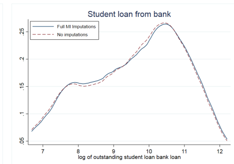
(b) Bank



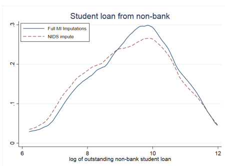
(c) Micro Loans



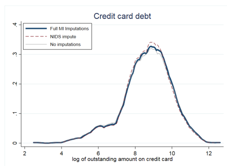
(d) Masho?



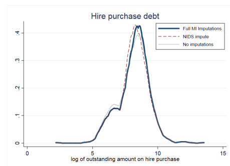
(e) Study



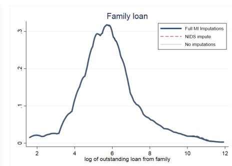
(f) Study Other



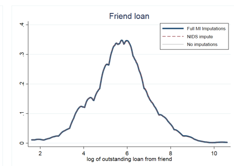
(g) Credit



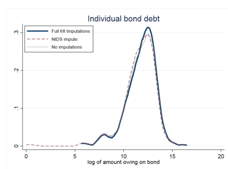
(h) Hire



(i) Family

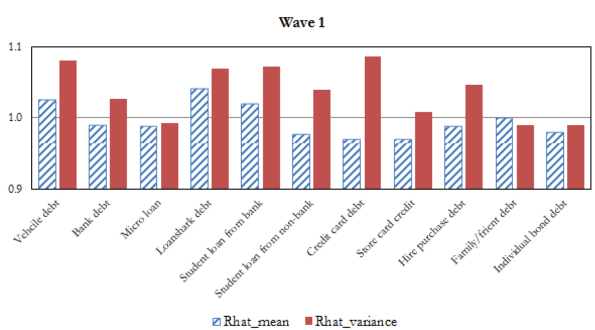


(j) Friend

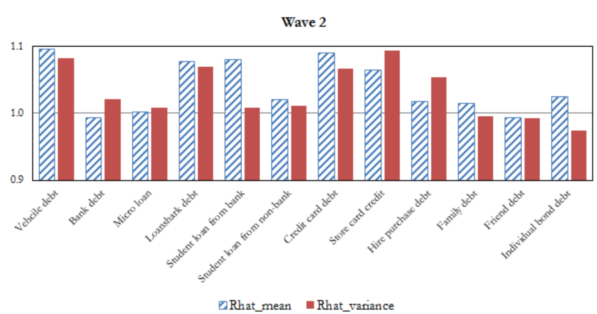


(k) Bond

Figure 12: Wave 4 imputations compared to NIDS imputations (when applicable)



(a) Wave 1



(b) Wave 2

Figure 13: Rhat for all imputed individual debt types

Table 2: Description of household income by wave

	Wave 1	Wave 2	Wave 3	Wave 4
Total number of observations	7 296	9 016	10 114	11 732
Income reponses	7 296	6 782	8 033	9 618
Responded	5 440	5 718	7 262	9 134
Non-response	1 856	1 064	771	484
Non-response, but provided bracket	474	651	493	325
% reported point value (if asked the question)	74.6	84.3	90.4	95.0
% reported point value plus bracket	81.1	93.9	96.5	98.3
Total percentage point increase in response rate	6.5	9.6	6.1	3.4
Upperbound	99.9th	99.9th	99.9th	99.9th
Value	R 88 975	R 229 122	R 236 850	R 106 055

Table 3: Comparing means from different imputation methods and no imputations for household income

Method used	Wave 1	Wave 2	Wave 3	Wave 4
No unfolding of brackets	R 3 094	R 4 025	R 5 067	R 10 680
Mid-point unfolding of brackets	R 3 394	R 4 030	R 5 089	R 10 667
Multiple imputation unfolding of brackets	R 3 299	R 4 004	R 5 056	R 10 639

Table 5: Number of respondents before and after imputations for households who responded that they have property debt, but did not provide a value

Method used	Wave 1	Wave 2	Wave 3	Wave 4
Original responses	313	264	226	244
Imputed responses	304	280	145	143
Total responses	617	544	371	387

Table 6: Comparing means from different imputation methods and no imputations for property debt

Method used	Wave 1	Wave 2	Wave 3	Wave 4
No imputations	R 237 328	R 286 174	R 315 265	R 368 651
MI imputed	R 198 728	R 241 570	R 292 066	R 323 632
NIDS imputed	None	None	None	R 279 827

Table 4: Debt variables available and imputed by wave

	Wave 1	Wave 2	Wave 3	Wave 4
Household property debt outstanding	x	x	x	x
Individual vehicle debt outstanding	x	x	x	x
Individual bank debt outstanding	x	x	x	x
Individual micro debt outstanding	x	x	x	x
Individual Mashonisa debt outstanding	x	x	x	x
Individual student loan from bank outstanding	x	x	x	x
Individual student loan from other outstanding	x	x	x	x
Individual credit card debt outstanding	x	x	x	x
Individual store card debt outstanding	x	x	x	x
Individual hire purchase debt outstanding	x	x	x	x
Individual loan from family or friend	x		x	
Individual loan from friend		x		x
Individual loan from family		x		x
Individual bond owing	x	x	x	x

Table 7: Responses to individual debt outstanding questions (Wave 1, Wave 3 and Wave 4), example of vehicle debt

Has vehicle finance (car payment)?	Number of observations
Refused	8
Missing	63
Yes	400
No	15 159
Total	15 630

Table 8: Responses to individual debt outstanding questions (Wave 2), example of vehicle debt

Has vehicle finance (car payment)?	Number of observations
Don't know	2
Refused	49
Missing	12
Not asked in Phase 2	733
Yes	182
Not asked in Phase 2	16 651
Total	17 629

Table 9: Number of responses before and after NIDS imputations and MI imputations

	Wave 1		Wave 2			Wave 3		Wave 4		
	Point values	Total including imputations (MI)	Point values	Total including imputations (NIDS)	Total including imputations (MI)	Point values	Total including imputations (MI)	Point values	Total including imputations (NIDS)	Total including imputations (MI)
Vehicle debt outstanding	223	400	140	181	191	176	259	361	435	435
Bank debt outstanding	351	523	316	413	421	592	814	1 225	1 411	1 411
Micro debt outstanding	50	81	41	41	51	45	59	196	211	211
Mashonisa debt outstanding	94	137	132	158	158	215	238	339	345	345
Student loan from bank outstanding	31	58	32	32	39	25	29	46	46	48
Student loan from other outstanding	17	36	22	22	30	27	39	60	60	76
Credit card debt outstanding	538	762	258	332	345	390	543	609	651	651
Store card debt outstanding	1 175	1 572	928	1 054	1 074	1 582	1 887	3 289	3 528	3 528
Hire purchase debt outstanding	338	457	249	296	296	489	634	807	907	907
Loan from family or friend	177	250				204	229			
Loan from friend			124	155	155			679	684	684
Loan from family			81	81	100			341	342	342
Bond owing	356	541		219	286	258	323		319	376
Total debt response rate	3 350	4 817	2 323	2 984	3 146	4 003	5 054	7 952	8 939	9 014
% increase with imputations		43.8		28.5	35.4		26.3		12.4	13.4

Table 10: Means of debt variables before and after NIDS and MI imputations

Vehicle debt outstanding	Wave 1	Wave 2	Wave 3	Wave 4
No imputations	R 73 521	R 124 841	R 93 508	R 146 464
NIDS imputations	None	R 124 858	None	R 131 153
MI	R 75 661	R 135 108	R 99 676	R 155 384
Bank debt outstanding				
No imputations	R 16 692	R 14 952	R 18 998	R 58 280
NIDS imputations	None	R 13 313	None	R 52 745
MI	R 17 902	R 15 587	R 19 981	R 56 695
Micro debt outstanding				
No imputations	R 4 419	R 19 838	R 6 299	R 6 471
NIDS imputations	None	None	None	R 6 508
MI	R 4 014	R 19 982	R 5 997	R 6 427
Mashonisa debt outstanding				
No imputations	R 1 811	R 989	R 1 252	R 1 482
NIDS imputations	None	R 942	None	R 1 468
MI	R 2 098	R 916	R 1 223	R 1 565
Student loan from bank outstanding				
No imputations	R 23 201	R 28 599	R 25 549	R 30 498
NIDS imputations	None	None	None	None
MI	R 22 639	R 28 193	R 26 070	R 30 920
Student loan from other outstanding				
No imputations	R 13 343	R 21 527	R 46 346	R 22 649
NIDS imputations	None	None	None	None
MI	R 15 426	R 23 901	R 47 319	R 23 573
Credit card debt outstanding				
No imputations	R 5 028	R 5 325	R 8 974	R 11 192
NIDS imputations	None	R 5 383	None	R 11 219
MI	R 5 997	R 5 628	R 9 948	R 11 283
Store card debt outstanding				
No imputations	R 1 930	R 2 345	R 2 814	R 3 317
NIDS imputations	None	R 2 333	None	R 3 216
MI	R 2 014	R 2 356	R 2 802	R 3 338
Hire purchase debt outstanding				
No imputations	R 3 986	R 4 880	R 5 416	R 7 374
NIDS imputations	None	R 4 566	None	R 7 020
MI	R 4 319	R 4 729	R 5 467	R 7 444
Loan from family or friend				
No imputations	R 4 642	None	R 1 405	None
NIDS imputations	None	None	None	None
MI	R 3 859	None	R 1 390	None
Loan from friend				
No imputations	None	R 567	None	R 737
NIDS imputations	None	R 502	None	R 736
MI	None	R 541	None	R 743
Loan from family				
No imputations	None	R 2 444	None	R 2 085
NIDS imputations	None	None	None	R 2 215
MI	None	R 2 197	None	R 2 081
Bond owing				
No imputations	R 184 909	R 232 048	R 224 795	R 323 170
NIDS imputations	None	None	None	None
MI	R 184 834	R 227 314	R 231 549	R 364 990