January 6, 2014

Imputing NCVS Income Data

Report

Prepared for

Bureau of Justice Statistics
U.S. Department of Justice
810 7th St NW
Washington, DC 20531

Prepared by

Marcus Berzofsky, DrPH
Darryl Creel, MS
Andrew Moore, MS
Hope Smiley-McDonald, PhD
Chris Krebs, PhD
RTI International
3040 E. Cornwallis Road
Research Triangle Park, NC 27709

RTI Project Number 0213170.001.002.001

RTI INTERNATIONAL

Imputing NCVS Income Data

Prepared by
Marcus Berzofsky, DrPH
Darryl Creel, MS
Andrew Moore, MS
Hope Smiley-McDonald, PhD
Chris Krebs, PhD

Any opinions and conclusions expressed herein are those of the author(s) and do not necessarily represent the views of the Bureau of Justice Statistics and the U.S. Department of Justice.

This document was prepared using Federal funds provided by the U.S. Department of Justice under Cooperative Agreement number 2011-NV-CX-K068. The BJS Project Managers were Lynn Langton, BJS Statistician, and Michael Planty, Victimization Unit Chief.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The National Crime Victimization Survey (NCVS) is one of the most important sources of information on criminal victimization in the United States. Each year, data are obtained from a nationally representative sample of about 40,000 households comprising nearly 75,000 persons on the frequency, characteristics, and consequences of criminal victimization. The survey enables the Bureau of Justice Statistics (BJS) to estimate the likelihood of experiencing rape, sexual assault, robbery, assault, theft, household burglary, and motor vehicle theft victimization for the population as a whole as well as for segments of the population.

Virtually all data collection efforts experience the challenge of missing data, and the NCVS is no exception. Because the NCVS is a panel data collection effort, with households in sample for 3 years and interviewed every 6 months, there are three possible types of nonresponse with the survey: unit, wave, and item. *Unit nonresponse* occurs when no information is collected from the sample member. *Wave nonresponse* is when the sample member responds to at least one wave of data collection but does not respond to at least one of the other waves of data collections. *Item nonresponse* is when the sample member responds to the wave but fails to provide information about one or more of the questions asked in the wave.

Although the NCVS, like most surveys, uses weight adjustments to account for unit and wave nonresponse, for most of the variables in the NCVS, nothing is currently done to address item nonresponse. The problem with these missing data is that, if not properly addressed, there is a loss of power and the potential that biased estimates will be produced from the respondent data. This issue is particularly problematic when it comes to the measurement of household income, which is one of the NCVS items that has historically suffered from high rates of nonresponse. The rate of missingness has increased over time, with the weighted proportion of household respondents that did not report a household income rising from 10% in 2000 to 32.4% in 2010.

The NCVS is not alone with respect to nonresponse for income, as income is generally known to have the highest rate of item nonresponse. Recent data from the British Crime Survey showed that 20% of the responses for income were missing or were otherwise inadequate for

reporting (Home Office, 2011). Income nonresponse rates for the 2011 American Community Survey (29% imputed; U.S. Census Bureau, 2011) were also comparable to the NCVS rates.

However, as BJS works to continually improve the utility of the NCVS for understanding changes in crime and its correlates over time, the high levels of missing income data cannot be ignored given established associations between victimization and socioeconomic factors, including income. Continuing to ignore the missing income data and drawing conclusions about the relationship between victimization and income solely from households that provided income information may lead to false interpretations of this relationship. Thus, the purpose of this report is to recommend possible imputation methods that can be used to impute income data for the NCVS longitudinal data. Although alternative methods for addressing missing data during analysis exist, imputation is a common approach used to address missing data issues "so that the resulting data set is, in a sense, complete. This is most convenient for large prospective databases that will be used for many different types of analyses by a number of researchers" (Engels & Diehr, 2003, p. 968).

The report first describes the problem of missing income data in the NCVS. It then details and assesses several potential approaches to imputing the missing data and, based on the findings, recommends the use of the single imputation hot deck method for the future imputation of NCVS household income data or any other variables for which there is item nonresponse.

SECTION 1. DATA IN THE NATIONAL CRIME VICTIMIZATION SURVEY

1.1 Measuring Income in the National Crime Victimization Survey

Within the NCVS, income is measured by a single categorical question which asks the respondent to choose from 14 different household income response categories. As shown in ***Table 1-1*** later in this section, the income ranges are not uniform across the fourteen categories. The household respondent is asked about income every other interview wave. Currently, in the interview waves in which the income question is not asked, a carry forward imputation method is used. The carry forward imputation assigns the reporting household income value to the current interview wave (e.g., if a household income level of 3 is reported in interview wave 5 then a level 3 is assigned as the household income in interview wave 6).

1.2 The National Crime Victimization Survey Data

The starting point for the preliminary analysis of missing income data was the 2006–2010 longitudinal file provided to RTI International as part of the NCVS Redesign Screening Questions project. This file includes only sample number 24, panels 1–6, and rotation groups 4–6. The purpose of using this cohort was to facilitate the determination of time-in-sample and allow for easier determination of patterns of missing data in the income question. The data presented in Tables 1-1 through 1-4 as well as Figures 1-1 and 1-2 are based on this cohort of respondents in the 2006–2010 longitudinal file. This file is being used for research purposes only, and the usual NCVS files are annual files that may only contain two interviews per household. Once the preliminary analysis was completed, the longitudinal file containing the cohort was discarded and the full annual files were used for imputation purposes.

The 2006–2010 longitudinal file format is at the person level, with one record per person per wave in sample. Each person is represented by seven records, regardless of whether there was wave nonresponse and whether or not the household had been replaced in the sample. The household remains in the sample even if the residents of the household change. Because income is measured at the household level, a household-level file was needed for the current analysis. To obtain this file, the 2006–2010 longitudinal file was first restricted to household respondents (hr_flag=1). This subset contained records for 29,297 unique households (i.e., a unique family

within a sampled address). These households come from 22,337 unique addresses, of which 17,373 addresses contained the same family for all seven waves. In other words, 4,964 addresses had two or more families residing at the address during the seven waves. For these households, the file was expanded to include seven records for each household with dummy records being inserted for household by time in sample combinations that did not include a household respondent as indicated by the hr_flag variable. These dummy records represented cases with wave nonresponse and did not distinguish between the different reasons for nonresponse (e.g., a household was no longer in the sample; a household had not yet entered the sample; a household was in the sample but did not respond or was not able to be contacted).

Once a file was created with seven records per household, the wave in which the first completed household interview had been conducted was identified. The income variable was then transposed based on the time since the first completed interview so that there was one record per household with the derived variable, "income_t1," representing the income response during the year/quarter of the first completed household interview (regardless of income item response status), "income_t2" representing income during the next year/quarter after the first completed household interview (regardless of wave/item response status), and so on through "income_t7." This method was chosen so that the income item response/nonresponse status was comparable across households. For example, suppose a replacement household (HH1) entered the sample during the fifth wave of data collection and was a wave respondent while a second household (HH2) entered the sample during the first wave and was a wave respondent for this interview wave. In this instance, the goal was to compare the income response from the first interview of HH1 (fifth wave for the given address) to the first interview of HH2 (first wave for the given address) rather than comparing the fifth wave from HH1 (which is actually their first time responding to the questionnaire) to the fifth wave of HH2.

As an example, consider the imputation of an ordinal income variable. From the NCVS-1 basic screen questionnaire (OMB No. 1121-0111, Form NCVS-1, Implementation Date 07-01-2008, covering July 2008 through December 2009), question 12a asks about household income. There are 14 possible response categories. Category 1 is the smallest income category and is for household incomes of less than $5,000. Category 14 is the largest income category and is for household incomes of $75,000 or more. The remaining 12 household income categories partition

the household incomes between $5,000 and $74,999. These categories are smaller at the lower end of household income and larger at the higher end of household income. ***Table 1-1*** shows how the categories are mapped to actual dollar amounts and the distribution of income over interview waves.

Table 1-1. Distribution of income over interview waves, 2006–2010

Income code	Dollar amount	Int. 1 count	Int. 2 count	Int. 3 count	Int. 4 count	Int. 5 count	Int. 6 count	Int. 7 count
Blank			7,794	12,047	14,346	15,955	17,366	18,670
1	< 5,000	1,002	541	258	178	145	119	109
2	5,000–7,499	576	360	214	151	112	97	89
3	7,500–9,999	579	433	288	245	188	156	146
4	10,000–12,499	813	546	323	261	271	230	187
5	12,500–14,999	596	407	259	220	196	170	188
6	15,000–17,499	586	428	275	226	221	193	194
7	17,500–19,999	669	500	320	254	238	209	171
8	20,000–24,999	1,457	1,059	739	607	529	464	443
9	25,000–29,999	1,210	893	634	563	495	446	448
10	30,000–34,999	1,370	1,054	777	654	587	511	406
11	35,000–39,999	1,218	964	650	600	548	492	451
12	40,000–49,999	1,986	1,577	1,176	1,023	942	838	724
13	50,000–74,999	3,275	2,769	1,974	1,773	1,644	1,498	1,325
14	75,000 and over	4,822	4,207	3,359	3,118	2,717	2,548	2,275
98	Item missing	9,138	5,765	6,004	5,078	4,509	3,960	3,471

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

In addition to comparing the income distribution across time in sample, knowing how the distribution of income among NCVS respondents compares to an external source can be informative for a couple reasons. First, the comparison informs whether the current income distribution is biased. Second, the comparison provides a benchmark for what the imputed distribution should look like. The American Community Survey (ACS) is a good external source for this comparison because it provides annual estimates of income using income categories similar to those of the NCVS. ***Table 1-2*** presents the income distribution among respondents in the NCVS compared to the distribution from the ACS. It appears that there is little bias incurred by the high level of item nonresponse in the NCVS and that the imputed distribution of income, if done properly, should not dramatically alter the current distribution.

Table 1-2. Estimated income distribution in the American Community Survey and respondents in the National Crime Victimization Survey, 2010

Income category	American Community Survey	Respondent
Less than $10,000	7.6	7.7
$10,000–$14,999	5.8	6.0
$15,000–$24,999	11.5	12.3
$25,000–$34,999	10.8	12.4
$35,000–$49,999	14.2	16.2
$50,000–$74,999	18.3	17.5
$75,000 or more	31.7	27.9

Sources: American Community Survey, 2010; Bureau of Justice Statistics, National Crime Victimization Survey, 2010.

Table 1-3 shows the level of missingness for the income variable at each interview wave, where missingness is defined as having an item missing code. The table presents the interview wave, which is the number of times the household has been interviewed, the number of households marked as item nonrespondents for the interview wave, and the percentage of households marked as item nonrespondents for the interview wave. The percentage is based on all the data for the interview wave and may not accurately reflect the true denominator because it does not exclude unit and wave nonrespondents for the interview wave (i.e., some of those who were item nonrespondents in earlier interview waves have shifted to be interview nonrespondents as time in sample increases). In addition, there was imputation of income in the even waves of data collection. These imputed values could not be distinguished from actual respondent values. Consequently, all the valid income values were treated as respondent values. Even so, the table provides some information about the minimum level of missingness for the income variables, which is considerable.

Table 1-3. Number of item nonrespondents and percentage of item nonresponse for income by interview wave, 2010

Interview wave	Number of item nonrespondents	Percent item nonresponse
1	9,138	31
2	5,765	20
3	6,004	20
4	5,078	17
5	4,509	15
6	3,960	14
7	3,471	12

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

Because households have likely been interviewed at least once and as many as seven times, the higher the number of interviews, the more missing data (unit nonresponse). Therefore, the percentage of data that falls into the missing category increases over time. Given the increase in missingness, even if the relative sizes of item respondents and item nonrespondents remain the same, the percentage of item nonresponse for an interview wave will decrease as the interview wave number increases because the denominator (i.e., the number of sampled households) decreases over time.

Table 1-4 shows the number of interview waves for which a household has an item nonresponse code for income across the seven interviews. It is not restricted to only those households that completed all seven interviews. The table shows the number of interview waves with an item nonresponse code, the number of households, and the percentage of all households. The interpretation of Table 1-4 is that 51% of households did not have an item missing code for income across the seven interviews, 17% of households had one income value with an item missing code across the seven interviews, 13% of households had two income values with an item missing code across the seven interviews, and so on. Typically, income is asked only during the first, third, fifth, and seventh interviews and values are carried forward from the first to the second, third to fourth, and fifth to sixth, which would typically prevent a household from having only one missing value for income. However, because it is possible for households to enter and leave the survey at different points in time and because unit and interview nonresponse were not counted in the item nonresponse code counts, it is possible to have any number of item missing values. For example, if a replacement household entered the survey during the seventh interview

period and the household was a unit respondent but did not respond to the income question, then that household would have one missing value for income.

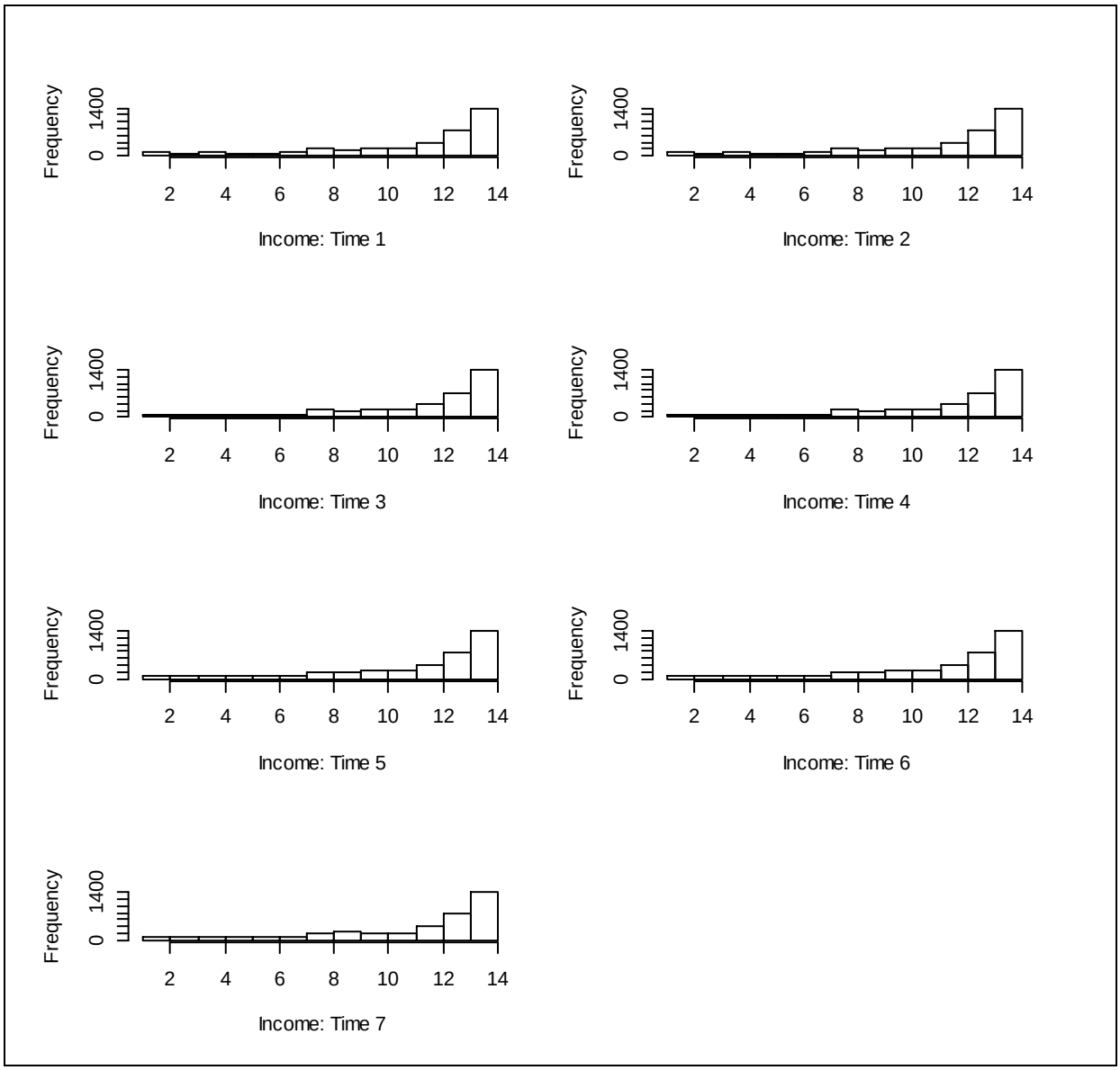
Table 1-4. Level of item nonresponse for income by household, 2010

Number of interview waves with item nonresponse code	Number of households	Percentage of households
0	14,932	51
1	4,855	17
2	3,762	13
3	1,869	6
4	1,402	5
5	1,231	4
6	546	2
7	700	2

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

The data shown in the following two figures are based on the 3,908 households that have valid values for income across all seven interview waves. First, in ***Figure 1-1***, the distribution of the ordinal income variable over the seven interview waves is similar. Note that the distribution is highly skewed to the right, and the most common income category is household income $75,000 and over. This pattern is consistent across interviews.

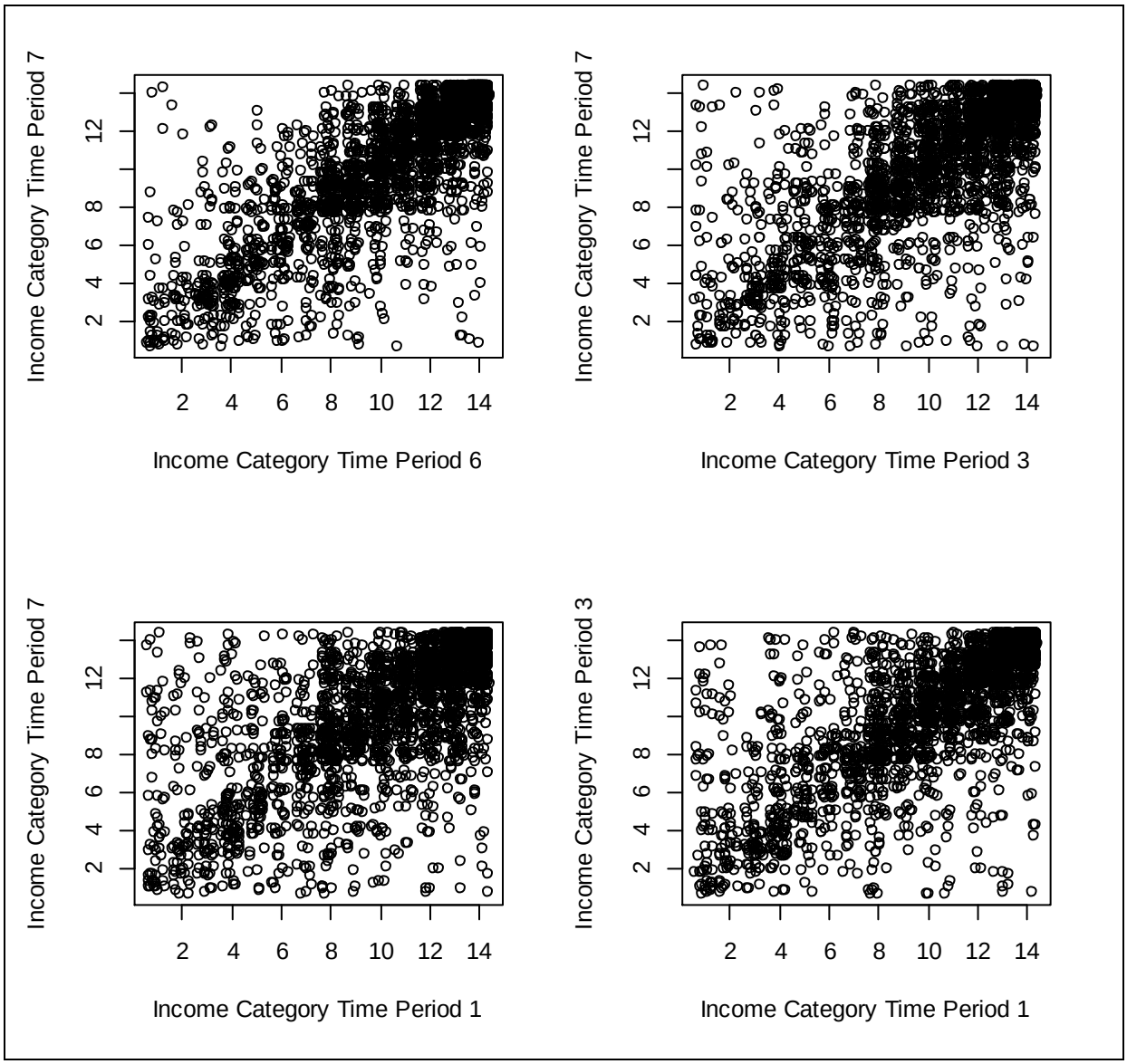
Figure 1-1. Distribution of income over time (i.e., waves), 2010



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

Second, ***Figure 1-2*** presents the relationship among these variables across the interview waves using four representative plots. The plots are the income categories for interview 7 by interview 6, interview 7 by interview 3, interview 7 by interview 1, and interview 3 by interview 1. No matter which interview waves are used, there is a strong relationship between the income categories over time. That is, households usually stay in the same income category over time. There is not much movement from one income category to another. As expected, the relationship is slightly attenuated as the interview waves move further apart.

Figure 1-2. Relationship of income over time



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

SECTION 2: IMPUTATION METHODS AND OPTIONS

2.1 Imputation Methods

The previous section described the relationship between income and item nonresponse and how that relationship changes across panel waves. This section describes the methods typically used to address item nonresponse.

Most imputation methods for longitudinal data treat all missing values as if they are from the same missing data mechanism—that is, *missing completely at random, missing at random,* or *missing not at random*. A rotating panel design will have planned missingness because of the changing panel membership and item missingness because of nonresponse. Thus two missing data mechanisms may be at work: (1) missing at random for the planned missingness and (2) missing not at random for the item nonrespondents. Recent research has started to look at this phenomenon. Giusti and Little (2011) have applied an imputation approach that treats missing data from two missing data mechanisms for the labor force survey of the Municipality of Florence in Italy. Their conclusion is that "income amounts are moderately sensitive" (p. 226) to deviations from the missing at random assumption commonly used by other imputation approaches. Consequently, this approach was not included in the imputation methods that were evaluated for several reasons: (1) the research is relatively new and requires more extensive investigation—for example, different data sets and Monte Carlo simulation; (2) no available commercial software exits to implement the methodology; and (3) Giusti and Little's results indicate only moderate sensitivity to the missing at random assumption.

Several studies have investigated imputation methods for panel data assuming the same missing data mechanism. These methods may use information from only the person missing the data value or from all the people in the data set. The methods may use information from only the wave in which the missing data value occurs or from all available waves. The two main categories of imputation methods for panel data are cross-sectional imputation methods and longitudinal imputation methods (Twisk & de Vente, 2002; Engels & Diehr, 2003), both of which are summarized below.

Cross-sectional imputation methods use the data from the wave in which the item nonresponse occurs. Examples of cross-sectional imputation include

a. mean (median) value imputation, where the mean is the average value of the variable for all respondents for the specific wave, which is calculated from the data overall or by specified respondent categories that could, for example, reflect sex, age, health status, or education;

b. hot deck imputation,[1] where the missing value is replaced by a valid value from a respondent for the specific wave; typically, the data are categorized and a donor is chosen from the category in which the missing value occurs; and

c. cross-sectional regression methods, where the predictor variables in the model are from the same wave in which the missing value occurs.

Longitudinal imputation methods use information from the wave in which the item nonresponse occurs along with information from other waves. Examples of longitudinal imputation include

a. last value carried forward, where the most recent reported value for a respondent is used for the missing value;

b. subsequent value carried backward, where the reported value for a respondent in a subsequent wave is used for the missing value;

c. mean (median) value imputation, where the average of the respondent's reported values is used as the imputed value;

d. linear interpolation, where the data pattern reflects the (1) reported value in a previous wave, (2) missing value in a wave, and (3) reported value in a subsequent wave and the imputed value is the average of the previously reported value and the subsequently reported value for a respondent;

[1]The term *hot deck* refers to the fact that the set of potential donors comes from the same data set. In contrast, *cold deck* imputation refers to the fact that the donors come from an external data set or source.

e. individual longitudinal regression, where the imputed values are derived from a regression that uses time in the model;

f. population longitudinal regression, where the imputed value is derived from a regression that uses previous values of the missing data, time, and additional predictors in the model; and

g. the "Little and Su" method (Little & Su, 1989), where the imputed value is based on a trend effect (column effect), an individual effect (row effect), and a residual effect, where the effects can be based on all respondent data or on data from respondents within specific categories.

Employing multiple imputation is an option that can be incorporated into several of the imputation methods (Rubin, 1987; Tang et al., 2005). Multiple imputation requires multiple imputed values for each missing value. For example, five imputations would require five imputed data sets, with each data set having imputed values for any missing data. The information from the five data sets is combined to produce the point estimates and standard errors that incorporate the uncertainty due to imputation.

2.2 Imputation Options

Although the focus for the NCVS is the ordinal income variable, the general approach of the following imputation options would be applicable to other types of variables as well (e.g., continuous, binary, or nominal categorical). Studies investigating the aforementioned longitudinal imputation methods include the following: (1) Twisk and de Vente (2002) assessed the impact of missing panel data on longitudinal data analysis, (2) Engels and Diehr (2003) compared the performance of several different imputation methods for missing panel data, (3) Tang et al. (2005) compared a few imputation methods for a longitudinal randomized clinical trial, and (4) Watson and Starick (2011) evaluated several imputation methods for different types of missing data in a household-based panel study. In general, findings from these studies suggest that when imputation methods are used with panel data, longitudinal imputation methods are generally preferred over cross-sectional methods, and multiple imputation may be preferred over single imputation because estimates of standard errors from multiple imputation more accurately

reflect the uncertainty due to imputation (Twisk & de Vente, 2002). Often, nonstatistical considerations influence the decision to use multiple imputation. For example, when multiple imputation is used, multiple public use data sets have to be created. Therefore, the future user of the data needs to be taken into account when selecting an imputation method, and the perceived difficulty of using the data may limit the types of imputation methods being considered.

The main advantage of the multiple imputation approach is that it produces standard errors that account for the uncertainty associated with the imputation process. The main disadvantage is that multiple data sets need to be created and used in the analysis, and some researchers may not be familiar with how to use the multiple data sets in their analyses. The advantage for the single imputation approach is that there is only one data set to be analyzed, and analysts can use appropriate data analyses techniques with which they are familiar. The disadvantage to the single imputation approach is that the single imputation approach does not reflect the uncertainty associated with imputed data.

Therefore, because single and multiple imputation techniques require different considerations, both were reviewed for the NCVS. The single imputation and multiple imputation techniques were compared in an effort to determine how best to impute household income in the NCVS.

2.2.1 Two Multiple Imputation Options

With an understanding of the level and pattern of missing income data in the NCVS, the distribution of the ordinal income variable, and the relationship among the income variables over the seven interview waves, two multiple imputation approaches to impute the income variables are recommended: (1) an explicit model approach and (2) a hot deck approach. Both approaches are longitudinal in nature and appropriate for use with panel data. That is, they would both use previous values of income as predictors in the models or in the creation of the imputation classes. Either of these approaches can be used to create a single imputation or multiple imputations. If any variables used in these approaches have missing values, these values would have to be imputed as well. The general imputation approaches are applicable to many types of variables, including income, which will be used as the example.

The explicit modeling approach would use an appropriate generalized linear model for the variable to be imputed. These models can be implemented in Imputation and Variance

Estimation Software (IVEware; Raghunathan, Lepkowski, Van Hoewyk, & Solenberger, 2001), which uses SAS macros. This imputation methodology would use sequential regression multivariate imputation and would proceed sequentially through each interview to impute the income variables. To impute income for a specific interview, income variables from the interview, whether imputed or not, and other specified variables (e.g., survey design, household characteristics, and variables related to income) would be used as predictor variables in the generalized linear models.

The hot deck approach would use bootstrap samples, nonparametric tree-based models to create imputation classes, and weighted sequential hot deck (WSHD) methods. For the number of imputations requested, a bootstrap sample (i.e., multiple independent with replacement samples) will be selected from the original data set, including respondent and missing values. Taking one of these bootstrap sample data sets as an example, each variable will be imputed using a prediction model based on tree-based methodology (i.e., recursive partitioning algorithm that divides the data space using binary splitting; Breiman, Friedman, Olshen, & Stone, 1984) to create imputation classes and WSHD (Cox, 1980) to select the imputed values. The imputation classes would be defined by making previous values of the variable to be imputed (i.e., income) from the interviews, whether imputed or not, and other specified variables (e.g., survey design, household characteristics, variables related to the variable to be imputed) available to the regression and classification tree algorithms. Once the imputation classes are created, WSHD will identify the imputed variables for the missing data.

Quality Checks. No matter which imputation method is used, quality checks should be conducted to ensure that the imputed income values are reasonable. These checks should include visualization of the respondent data, imputed data, combined data (i.e., both the respondent and imputed data combined), and distributions across the multiple imputations for each income variable. The quality checks should also investigate the bivariate relationships between all the imputed income variables to check the patterns of the respondent data and the imputed data for each pair of income variables. Finally, relationships between the income variables and other relevant variables (e.g., survey design, household characteristics, and variables related to the variable to be imputed) should be evaluated.

2.2.2 A Single Imputation Option

If multiple imputation is not feasible, then a single imputation process for the NCVS variables to be imputed should be considered. Several such methods are worthy of deliberation.

Single Imputation Linear Model Approach. The explicit modeling approach would use an appropriate generalized linear model for the variable to be imputed (i.e., income) as in the multiple imputation approach, but, instead of creating multiple imputations, it would create only a single imputation.

Single Imputation Hot Deck Approach. The general single imputation hot deck (SI-HD) methodology is a little different from the multiple imputation hot deck (MI-HD) approach and consists of three steps. The first step, if applicable, is logical or deterministic imputation. That is, if the imputed value can be deduced from the logical relationships with other variables, then that information is used to deterministically impute the value for the recipient. The second step, the first part of the random, or stochastic, imputation process, uses a tree-based methodology to create imputation classes. The third step, the second part of the random imputation process, uses hot deck imputation to select the imputed values. Specifically, for the random imputation, a relatively homogenous group of observations is identified using the tree-based prediction model, and independently within these groups, a random donor's value is selected to impute a reasonable value for the recipient using hot deck.

Quality checks, which are emphasized throughout both the linear model and hot deck single imputation approaches, are used to ensure that the data are consistent and to construct the imputation specifications that would accompany the data. The imputation specifications include the variable name, variable label, skip patterns, values that need to be imputed, valid values, related variables, and special instructions. The special instructions usually identify logical relationships that must be maintained between the variable to be imputed (i.e., income) and the other variables.

If the first round of imputation quality checks identifies inconsistencies in the data received for these variables, these inconsistencies can be replicated in the imputation process. Because the imputation process uses the valid responses for the NCVS data set as imputed values for the nonrespondents, it is possible to ensure that each variable to be imputed adheres to the

NCVS imputation specifications for skip patterns, valid values, and logical relationships. After reconciliation of any inconsistencies, the actual imputation process can begin.

Both the linear model and hot deck single imputation approaches are designed to impute all missing data as effectively and efficiently as possible using valid donor information, such that the process can be completed within a very short time frame after the end of data collection and still maintain the desired quality. The aim is to replace missing data with data that are reasonable and valid values for all cases.

Variables requiring imputation (i.e., the main variable of interest, household income, as well as all other variables used to inform the imputation process) are imputed sequentially. However, some variables that are related substantively or have similar patterns of missingness will be grouped together into blocks, and the variables within a block can be imputed simultaneously. The order in which variables, or blocks of variables, are imputed is based primarily on the level of missing data. The variables with lower levels of missing data are imputed before the variables with higher levels of missing data. All previously imputed variables will be available to be used when imputing additional variables.

When a variable is selected for imputation on the basis of its level of missing data, three specific pieces of information are evaluated. First, logical consistency is checked to make sure that any known relationships are maintained throughout the imputation process. Second, the pattern of missing data is evaluated to determine whether other variables should be included to create a block of variables requiring imputation. Finally, the imputation class variables and sorting variables are identified.

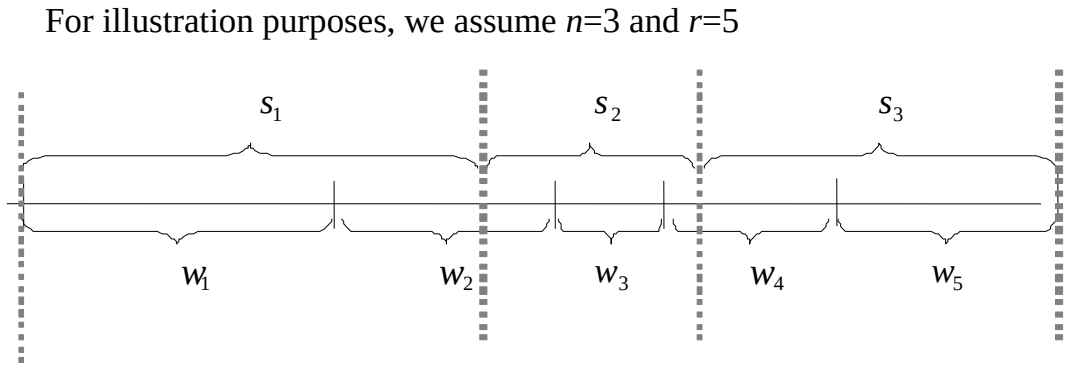
All stochastic imputations use a tree-based methodology to create imputation classes and the WSHD methodology (Cox, 1980; Iannacchione, 1982; Research Triangle Institute, 2012) within imputation classes. The imputation classes are formed using nonparametric classification trees (Breiman et al., 1984). The nonparametric classification trees form imputation classes from a prediction model based on the observations with valid values for the variable requiring imputation. The nonparametric classification tree recursively splits the cases into homogenous groups, which are used to define the imputation classes. The observations with missing values

for the variable to be imputed are assigned their imputation class on the basis of the same variables used in the tree splits.

The WSHD methodology replaces missing data with valid data from a donor record within an imputation class. The WSHD methodology also incorporates sorting within imputation class for additional control and uses the sample weight of each record in the donor selection process. The imputation classes in the application of the WSHD methodology are formed by identifying variables related to the variable requiring imputation. Data are sorted within each imputation class to increase the chance of obtaining a close match between donor and recipient. Within each imputation class, the hot deck process is sequential because the order of the recipients and donors in the file matters, and weighted because the weights are used to determine the relative sizes of donors and recipients. Recipient weights are scaled to the same overall size of the donor weights. The recipient and donor weights are aligned parallel to each other. The recipient weights divide the donors into zones for each recipient. A donor is randomly selected on the basis of the relative sizes of the donor weights in the zone for the recipient.

For ***Figure 2-1***, r is the number of item respondents (5), w_h is the sample weight for the hth respondent, n is the number of item nonrespondents (3), and s_i is the scaled weight for the ith nonrespondent. The zones are separated by the vertical dotted lines, which are based on the size of the scaled nonrespondent weights. That is, the scaled weights for the nonrespondents create the zones from which a donor will be selected. For example, in zone 1 for the first nonrespondent, the first or second respondent can be selected as a donor. In zone 2 for the second nonrespondent, the second, third, or fourth respondent can be selected as a donor. Finally, in zone 3 for the third nonrespondent, the fourth or fifth respondent can be selected as a donor. The selection probabilities within the zones are proportional to the size of the respondent weights within the zone.

Figure 2-1. Weighted sequential hot deck algorithm



Note: r is the number of item respondents (5), w_h is the sample weight for the hth respondent, n is the number of item nonrespondents (3), and s_i is the scaled weight for the ith nonrespondent. Zones, separated by the vertical dotted lines, are based on the size of the scaled nonrespondent weights. That is, the scaled weights for the nonrespondents create the zones from which a donor will be selected. The selection probabilities within the zones are proportional to the size of the respondent weights within the zone.

Once the donors have been selected, the imputation process is complete, and the second round of quality checks is conducted by the statistical programmer who has implemented the imputation specification.

The third round of quality checks is conducted by a statistician not involved in the actual imputation process. This round of quality checks includes the imputation diagnostics and consists of four quality checks: number of times a donor is used, overall imputation checks, imputation checks by class variables, and multivariate consistency checks. The check for the number of times a donor is used ensures that donors were used a reasonable number of times, as using a donor too many times may indicate that an imputation class has too few donors and the class needs to be enlarged. There are no definitive rules for the number of times a donor is used. Generally, the level of missingness is examined to get a general sense of how many times the donors will be used. For example, when a variable has 33% missing and 67% valid values, the expected number of times a donor would be used is about one-half. Here, one can think loosely of a Poisson distribution. Note that reviewing the number of times a donor is used is done only with the SI-HD approach. It is a basic check to ensure that nothing has gone seriously wrong (e.g., expecting to see donors used one time or a few times, but seeing that a single donor was used hundreds or thousands of times). The overall imputation checks compare the distributions, weighted and unweighted, for each level of the imputed variable before and after imputation. Differences of 5% or more are flagged and examined to see if changes should be made to the

imputation specifications. The imputation checks by class variables compare the distributions, weighted and unweighted, for each level of the imputed variable in the defined imputation classes before and after the imputation. Differences of 5% or more are flagged for further review. Finally, multivariate consistency checks ensure that relationships between variables are maintained and that any special instructions for the imputation are implemented properly.

If any of the four aforementioned diagnostic checks indicate a problem—that is, overuse of a donor, substantial deviations from the weighted sums, or any identified inconsistencies—the imputation process is evaluated and, if necessary, revised and rerun.

2.3 Cycling

Cycling is an integral part of the multiple imputation process, as it basically helps reinforce the relationships among variables, but cycling also benefits the single imputation process. Thus, if a single imputation process is adopted, adding cycling to the process is recommended, and the cyclic n-partition hot deck (Marker, Judkins, & Winglee, 2002)[2] is a preferred method, as discussed in Judkins (1997). It involves iteratively cycling through n-partition hot decks. The first iteration follows the current single imputation process, which uses the complete response variables and any previously imputed variables. On subsequent iterations, all variables on the data set would be available for the tree-based methodology to create imputation classes. Otherwise, the multiple imputation process is the same as the single imputation process applied to multiple data sets. The general idea for this approach is Bayesian. As Marker et al. (2002, p. 334) elaborate:

> This method was inspired by Bayesian methods but retains the semiparametric features of the hot deck. No strong assumptions are required about distribution shapes or about prior distributions for parameters. Instead, deliberate choices are made about which features of the covariance structure deserve the best preservation efforts.

This cycling approach is similar to the University of Michigan's IVEware software (Raghunathan et al., 2001) iterative procedure and could be done with or without multiple

[2] David Judkins is currently (2012 Joint Statistical Meetings) referring to this as *p*-cyclic partition hot deck. He changed from *n* to *p* because *n* is often used to denote the number of observations and *p* the number of variables.

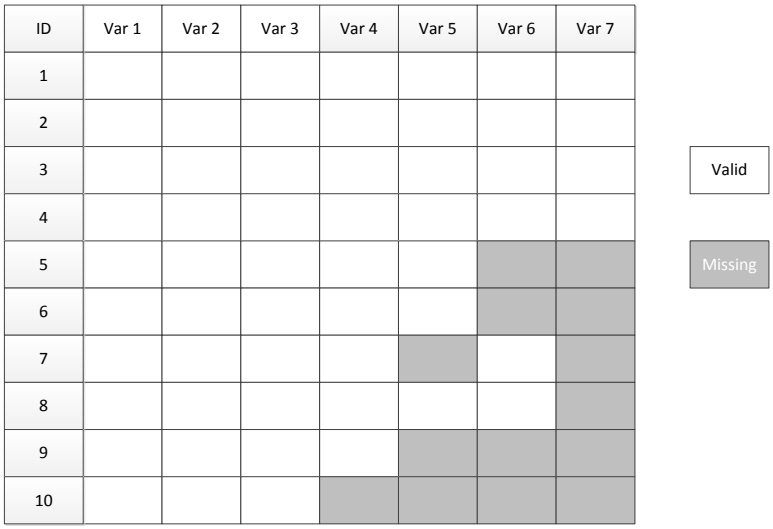
imputation. The goal is to have the imputed values converge to appropriate reasonable values. The common number of cycles to use is five.

This section describes the cycling step in detail, primarily relying on figures. There are essentially two phases required for cycling. The first phase is to create the initial imputed values (i.e., the aforementioned single imputation process) and the second phase is to conduct the actual cycling process, which may be iterated several times.

2.3.1 Phase One: Initial Imputed Values (Single Imputation Process)

In this example data set structure (***Figure 2-2***), the rows represent observations and the columns represent variables. The white rectangles contain valid responses, and the light gray rectangles contain missing responses.

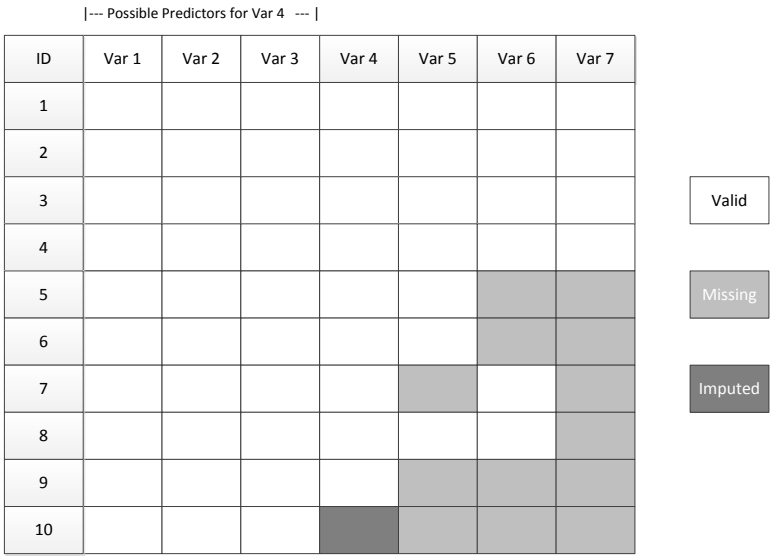
Figure 2-2. Initial state of missingness in data by variable



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

This begins the initial imputation process. In ***Figure 2-3***, the variable with the least amount of missingness, variable 4, is imputed first. The possible predictors are variables 1, 2, and 3—that is, the complete response variables. This creates the initial imputed value for variable 4. The dark gray rectangle contains the initial imputed value.

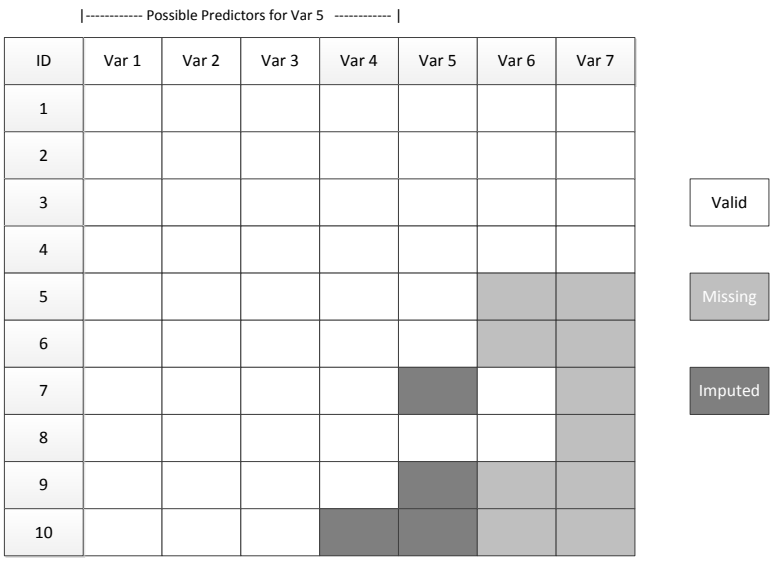
Figure 2-3. Initial imputation for variable 4



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

This is the second step in the initial imputation process. In ***Figure 2-4***, the variable with the second least amount of missingness, variable 5, is imputed. The possible predictors are variables 1, 2, 3, and 4—that is, the complete response variables and the previously imputed variable 4. This creates the initial imputed values for variable 5. The dark gray rectangles are the initial imputed values.

Figure 2-4. Initial imputation of variable 5



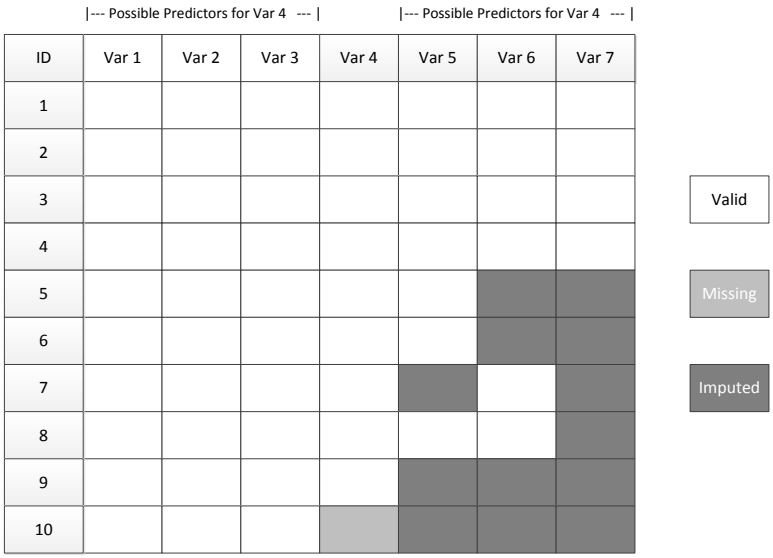
Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

This process continues until all variables with missing data have been imputed. Here, the figures for the initial imputation for variables 6 and 7 are not shown.

2.3.2 Phase Two: Cycling to Update Imputed Values

This is the first step in the cycling process. In ***Figure 2-5***, the variable with the least amount of missingness, variable 4, has the missing values recreated (the light gray rectangle). The possible predictors are variables 1, 2, 3, 5, 6, and 7—that is, the complete response variables and all the other imputed variables.

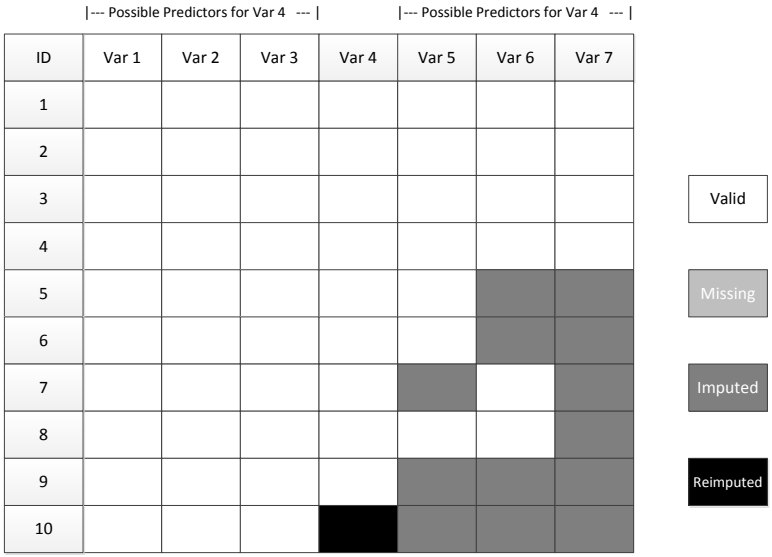
Figure 2-5. Recreate missingness for variable 4



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

Figure 2-6 shows that the missing value for variable 4 has been reimputed (see black rectangle). This process continues until all variables with imputed values have been reimputed. Here, the figures for re-creation of the missingness and reimputation for variables 5, 6, and 7 are not shown.

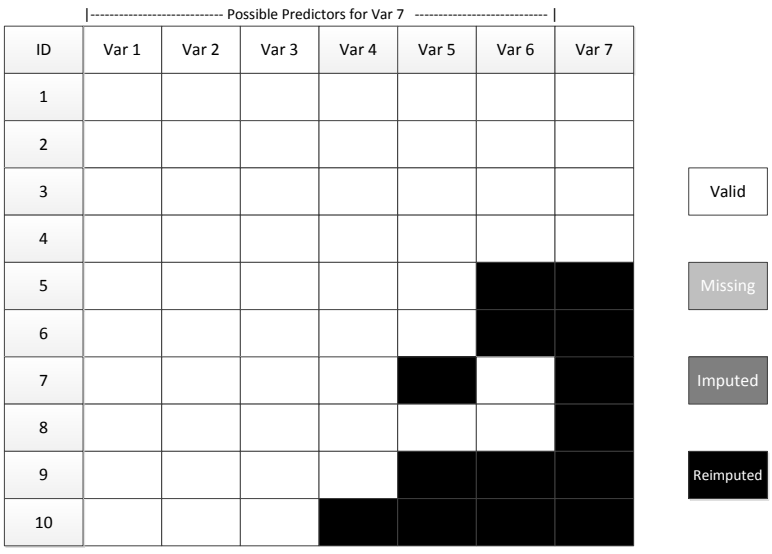
Figure 2-6. Reimputation of variable 4



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

Figure 2-7 shows that the missing values for variable 7 have been reimputed. Now that all the variables have been reimputed, one cycle has been completed, yielding a rectangular data set with no missing values. The cycling process may be repeated multiple times. That is, it is possible, and often preferable, to re-create the missingness for variable 4, reimpute variable 4, and continue through variables 5, 6, and 7 to complete a second cycle. Usually, several cycles are used.

Figure 2-7. Twelfth cycle of imputation



Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file.

SECTION 3: INVESTIGATING RESPONDENT DATA, SINGLE IMPUTATION, AND MULTIPLE IMPUTATION OF THE INCOME VARIABLES IN THE 2010 NATIONAL CRIME VICTIMIZATION SURVEY

3.1 Different Imputation Techniques Being Considered

The two general imputation approaches considered as part of this investigation are the hot deck method and the linear model method. Both of these approaches can be implemented in a single or multiple imputation process and in a cross-sectional or longitudinal manner. Because the analysis of the pattern of missing income responses in the NCVS shows that each household is likely to have at least one response to household income during the seven interview waves, the analysis was restricted to longitudinal approaches. However, among hot deck and linear model methods and how they are implemented (single imputation or multiple imputations), it is not clear which is preferable for the NCVS. Therefore, the analysis considered all four possible methods because they each offer advantages and disadvantages for the NCVS. Each method is briefly described below and some of their key advantages and disadvantages are identified in ***Table 3-1.***

- SI-HD: For the variable to be imputed, the SI-HD approach randomly selects a donor from the data set of respondents and uses the donor's value to replace the missing value for the recipient.

- (MI-HD: Repeats the process conducted in SI-HD a preset number of times.

- Single imputation linear model (SI-LM): A model-based approach to estimate the missing value for the recipient using a set of auxiliary variables to predict the imputed value.[3]

- Multiple imputation linear model (MI-LM): A model-based approach that repeats the process conducted under a SI-LM a preset number of times.

[3] IVEware, which was used for the linear model methodology, does not have a cumulative logistic regression model that would be used to model an ordinal outcome variable. Therefore, users must choose to treat the ordinal outcome variable either as a nominal categorical variable or as a continuous variable. The authors chose to treat the ordinal outcome variable as a continuous variable and round the results to an integer.

Table 3-1. Advantages and disadvantages of imputation approaches

	Advantages				
Imputation approach	**Ease of implementation**	**Always produces valid categorical value**	**Single version of data set required for analysis**	**Can incorporate many auxiliary variables**	**Accounts for variability in the imputation process**
SI-HD	X	X	X		
MI-HD		X			X
SI-LM	X		X	X	
MI-LM				X	X
	Disadvantages				
Imputation approach	**Analytically more cumbersome**	**May produce non-integer values requiring rounding**	**Multiple versions of the data needed for analysis**	**Limited number of auxiliary variables can be used due to donor pool size constraints**	**Does not account for the variability of the imputation process**
SI-HD				X	X
MI-HD	X		X	X	
SI-LM		X			X
MI-LM	X	X	X		

Note: HD, hot deck; LM, linear model; MI, multiple imputation; SI, single imputation.

3.2 Analytic Objectives

To determine the most appropriate imputation process for income in the NCVS, each of the four imputation procedures were assessed by three criteria:

1. Consistency of point estimates: how consistent is the distribution of the imputed values with the respondent data?

2. Variability of the imputations: how much variation in the imputed values does the imputation procedure create?

3. Usability and ease of implementation by the user: how easily can a user use the imputed data in an analysis?

3.2.1 Methodological Approach

In an effort to determine which imputation approach best meets the needs of the NCVS, each method was implemented on the 2010 NCVS public use data file. The longitudinal nature

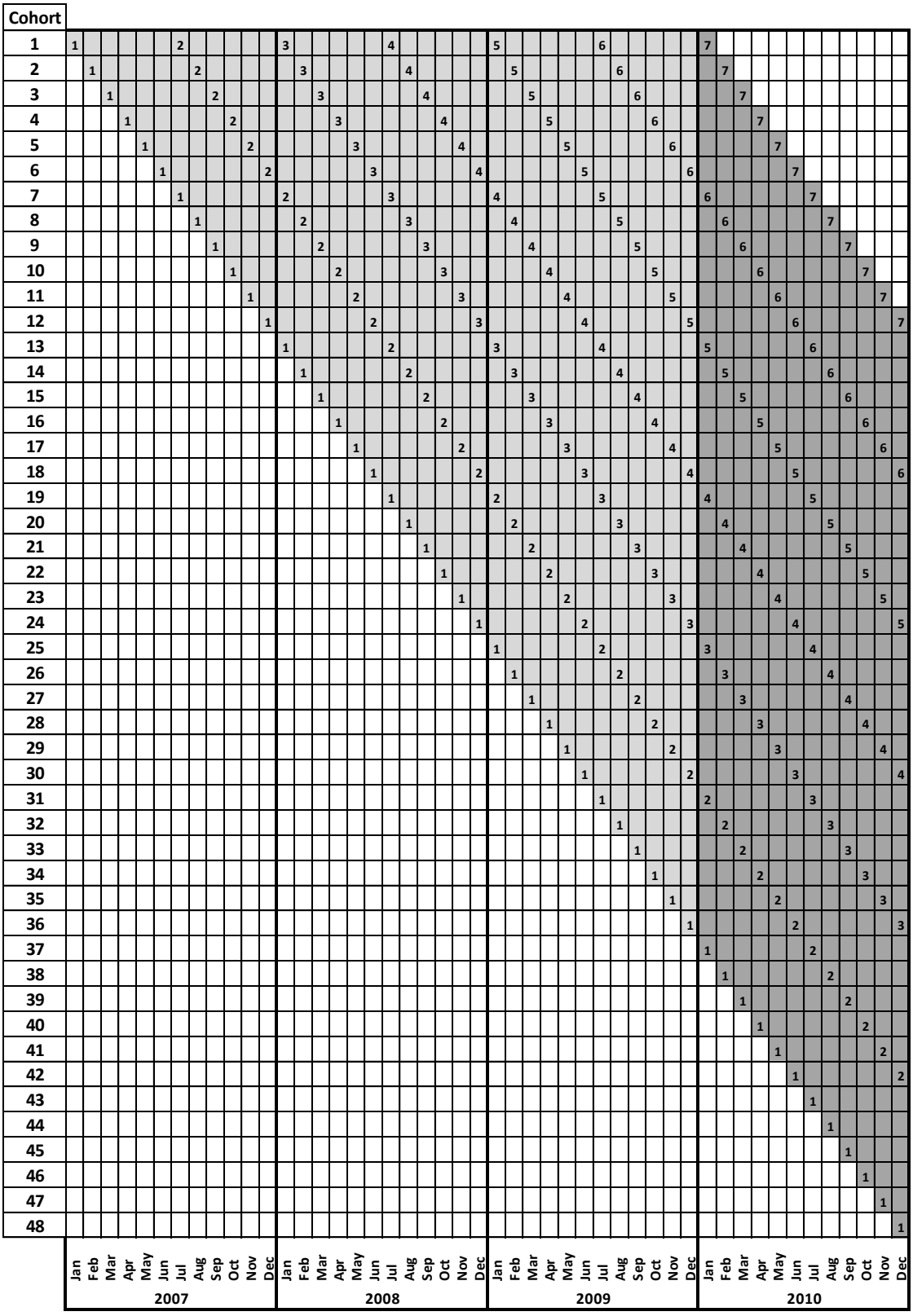
of the NCVS design (i.e., rotating panel design where panel members are in the survey for 7 waves over a 3.5-year period) allows researchers to take advantage of prior responses in the imputation process. To do this, the data files from 2007 to 2009 needed to be incorporated to derive benefit from earlier interviews.

3.2.2 Preparation of Data

To implement the proposed approach for applying income data for respondents interviewed in 2010, the first step in the data preparation process was to combine the data from 2007, 2008, 2009, and 2010. Even though the NCVS is a panel survey, for publication purposes it is treated like a series of cross-sectional surveys for each calendar year. That is, the annual public use files do not incorporate information from a respondent's prior waves that occurred in prior calendar years. However, given the structure of the current data file, it was possible to take advantage of the data from all of a respondent's interviews to impute income in 2010 even if interviews occurred before 2010. Therefore, a data set was constructed that linked the respondents in 2010 to all of their prior waves. Because a respondent is in the NCVS panel for 3.5 years (interviewed every 6 months, a total of seven times), a respondent who was in their seventh wave in 2010 had their first wave in 2007. Therefore, it was necessary to link the annual cross-sectional files from 2007, 2008, and 2009 to the 2010 data to ensure that all possible waves were present for respondents who were interviewed in 2010. However, these earlier years of data included respondents who completed their seventh wave before 2010 (i.e., they do not appear on the 2010 file). These respondents were extraneous for the analysis, and therefore were removed from the panel data set. Thus, only respondents whose seventh interview occurred during 2010 were included in the final analysis file.

Figure 3-1 provides an overview of the combined data set and illustrates how cohorts of respondents rotate in and out of the sample. The months and years in which relevant NCVS interviews occurred appear at the bottom of the figure. The interview waves are identified with corresponding numbers (1–7) in the figure. The interview waves that occurred in 2010, for which income data are imputed, are shaded dark. The interview waves that occurred before 2010, which are supplying the data used to impute income for interviews that occurred in 2010, are shaded more lightly.

Figure 3-1. Design of the National Crime Victimization Survey: 7 interview waves, 6 months apart, over time by cohort, 2007–2010



Respondents who completed their wave 1 interviews in January 2007 were on schedule to complete their wave 7 interviews in January 2010; therefore, for the purpose of trying to impute their income data for their wave 7 interviews in 2010 (January), it was possible to use or draw from up to six previous waves of interview data, going back to 2007. Similarly, respondents who completed their wave 1 interviews in June 2010 were on schedule to complete their wave 2 interviews in December 2010; therefore, for the purpose of trying to impute their income data for their wave 2 interviews in 2010 (December), it was possible to use or draw from only one previous wave of interview data, wave 1 in June 2010. The data and the staircase being used to present them begin in 2007 and extend to the years 2008, 2009, and 2010. Finally, one subset of respondents is in every year from 2007 to 2010, and some respondents in 2010 do not have any information from earlier years. The only data included in the analysis data set were those from respondents who were interviewed in 2010.

To combine the data sets across years and implement the imputation procedures, the data files were transposed from a hierarchical file (as is found in the public use files), with multiple records per respondent by interview wave, to a single record file, with one record per respondent and variables renamed to indicate from which wave they came. Once the data were transposed, the 4 years of data were merged by household number to identify when new families moved into a household. This process resulted in a data set with up to seven observations for income per household record. Household income is only asked every other interview. In the interviews in which household income is not asked about, the Census Bureau employs a carry-forward imputation approach, inserting the previously reported income as the household's response. Ideally, the process would delete these carry-forwarded income values and impute them using the proposed approaches. However, when an interview in which income is supposed to be asked about is missed, income is asked about in the following interview. No indicator on the public use file indicates whether the income value was reported or carried forward. Therefore, it is not possible to determine with certainty if the income value in the public use file came directly from a respondent or was carried forward; thus all nonmissing income values were treated as reported. A household could have fewer than seven income values due to nonresponse. Fewer than seven income values could also occur if a new family moved into a sampled household or if a household's seventh interview was scheduled to occur after 2010. For purposes of this analysis,

these households were excluded from the analyses. Instances of missing income data in the household remaining in the analysis data set were flagged for imputation.

3.2.3 Determining Key Covariates Through Tree Analysis

A tree analysis was conducted to determine which variables were best suited as predictors of a household's missing income level that could be used in the creation of imputation classes. The tree package in R was used to identify the key predictors (http://cran.r-project.org/web/packages/tree/tree.pdf). Thirty-seven variables were identified as possible predictors of a household's income level. These variables included various household characteristics as well as characteristics of the principal person in the household (usually the person considered the head of the household). ***Table 3-2*** lists the variables considered as predictors for use in creating the imputation classes.

The R tree package uses a recursive partitioning approach that looks at the amount of deviation in a group (or node) to determine if a further split should be created. In other words, if the addition of a particular variable does not decrease the amount of variability among the observations in a group by a specified amount, then a new node is not created. For the analysis, the minimum branch node size (i.e., the number of observations in any one level within a node) was set to 100 observations and the minimum size of a node was set to 200.

After implementing the tree package on the 37 potential predictor variables it was determined that a household's income in prior year and quarters (V2026 by year and quarter) best predicted a household's income when missing. In other words, none of the other 36 variables reduced the variability (i.e., created more homogeneous groups) enough to merit being included as variables in the analyses used to create the imputation classes.

Table 3-2. Possible predictor variables used to determine imputation classes

V2013: Special Place/Group Quarters (1=Yes, 0=No)	V2040A: Principal Person Race (1=White, 2=Black, 3=Amer Ind, 4=Asian, 5=Haw/Pac Isl, 6=Multiple Races)
V2015: Household Tenure	V2041: Principal Person Hispanic Origin (1=Hispanic, 0=Non-Hispanic)
V2017: Land Use (1=Urban, 0=Rural)	V2071: Number of Household Members 12 Years and Older
V2021: Type of Living Quarters	V2072: Number of Household Members Younger than 12 Years
V2024: Number of Housing Units in Structure	V2074: Operate Business from Address (1=Yes, 0=No)
V2025: Direct Outside Access (1=Yes, 0=No)	V2078: Number of Vehicles Owned (1=1, 2=2, 3=3, 4=4 or More)
V2025A: Gated or Walled Community (1=Yes, 0=No)	V2119: College or University (1=Yes, 0=No)
V2025B: Building with Restricted Access (1=Yes, 0=No)	V2120: Public Housing (1=Yes, 0=No)
V2026: Household Income	V2126B: Place Size Code
V2033: Principal Person Age	V2127B: Region (1=NE, 2=MW, 3=S, 4=W)
V2034: Principal Person Current Marital Status	V2129: CBSA MSA Status (1=Central City of MSA, 2=MSA But Not Central City, 3=Not MSA)
V2035: Principal Person Marital Status Last Survey Period	V2132: Principal Person Attending School (0=Reg School, 1=College, 2=Trade, 3=Vocational, 4=None of the Above)
V2036: Principal Person Sex	Number in Household That Worked in Last Week
V2037: Principal Person in Armed Forces (1=Yes, 0=No)	Number in Household That Worked in Last 6 Months
V2038: Principal Person Education (1=LT HS, 2=HS Grad, 3=Some Coll, 4=Assoc Deg, 5=Bach Deg, 6=Mast Deg, 7=Prof Deg, 8=Doct Deg)	Highest Educational Attainment in Household

Source: Bureau of Justice Statistics, National Crime Victimization Survey (NCVS), 2010 NCVS-1 basic screen questionnaire.

3.3 Final Data Set Preparations

Once the best predictor variables were determined, a data set containing only those variables and the income variables for 2010 was constructed. This data set contained eight variables: the two income values from 2007, 2008, 2009, and 2010 for each household.

In an effort to effectively use the income variables from 2007, 2008, and 2009 to impute the income variables in 2010, any of the 6 income variables from 2007-2009 that were missing prior to imputing the income variables from 2010 were imputed. This includes the income values that were carried forward in the waves in which the household was not asked about its income. These income variables were imputed in a two-step process based on the number of income values reported by a household respondent. The imputation procedure used was the same as the imputation procedure to be applied to the 2010 data (i.e., two data sets were created, one using hot deck imputation and one using linear modeling imputation). The first step, for respondents who had at least one reported income, was to initially impute across the interview waves for which the respondent did not have a reported income by using the mean income value for the respondent from other interview waves. This is known as *imputing horizontally*, across the file, so that only the reported income values for that specific respondent are used to impute the missing income values for the respondent. The income may have been missing because it was not reported by a respondent for the individual question (i.e., item nonresponse) or because of wave nonresponse.

The second step was to impute missing income data in a specified interview wave randomly, using the income distribution for that same interview wave. This is known as *imputing vertically*, up and down the file, so that the reported income distribution for a specific interview wave was used to impute the missing income values for that specific interview wave. The imputation process was random imputation based on the relative sizes of the respondent income categories. That is, the probability of imputing a specific income value was proportional to the income values size in the distribution of incomes based on the respondent data.

3.4 Imputation Procedures

For both hot deck imputations and linear modeling imputations, the single imputation and multiple imputation data sets were created simultaneously. Generally, as the level of missingness increased, the amount of variability in the imputation process increased. Therefore, the number of times an item was imputed, in a multiple imputation setting, may need to be increased to account for that increased variability. The income variables in the 2010 NCVS have relatively high levels of missingness, between 32% and 33% unweighted. Consequently, for the multiple

imputation procedures, to assess the level of variability in the imputation process, several different values for the number of multiple imputations were used, including 5, 10, 15, 20, and 25.

For each of the imputation procedures, the input imputation data set was run through the cycling process 25 times (see ***Section 2.5*** for details on cycling). The single imputation data set was the first of the imputed data sets. For the 5-imputation MI data set, the first 5 data sets were used. Similarly, for the 10-, 15-, and 20-imputation MI data sets, the first 10, 15, and 20 data sets, respectively, were used. For the 25-imputation MI data set, all 25 imputed data sets were used.

3.4.1 Hot Deck Procedures

For the hot deck imputation approach, bootstrap samples were taken from the original imputation data set to create the 25 data sets. The 25 bootstrap data sets were independently run through the cycling process to create the final imputations. SUDAAN's HOTDECK procedure uses WSHD methodology (Cox, 1980).

3.4.2 Linear Model Procedures

For the linear model imputations, 25 imputation data sets were created using the IVEware SAS macros (Raghunathan et al., 2001). Then the appropriate number of data sets was used to create the single imputation and the multiple imputation estimates. The conditional regression is based on a normal linear model. Each imputation consists of c "rounds." Round 1 begins by regressing the variable with the fewest number of missing values, Y1 on X, imputing the missing values under the appropriate regression model. Assuming a flat prior (i.e., a constant value that does not favor any particular parameter value over another) for the regression coefficients, the imputations for the missing values in Y1 are the draws from the corresponding posterior predictive distribution. The following procedures would then be implemented:

1. Update X by appending Y1 appropriately (for example, dummy variables, if it is categorical) and move on to the next variable, Y2, with the next fewest missing values.

2. Repeat the imputation process using updated X as predictors until all the variables have been imputed. That is, Y1 is regressed on U = X ; Y2 is regressed on U = (X, Y1) where Y1 has imputed values; Y3 is regressed on U = (X, Y1, Y2) where Y1 and Y2 have imputed values; and so on.

3. Repeat the imputation process in rounds 2 through c, modifying the predictor set to include all Y variables except the one used as the dependent variable. Thus, regress Y1 on X and Y2, Y3, ..., Yk; regress Y2 on X and Y1, Y3, ..., Yk; and so on.

As described above, repeated cycles continue for a prespecified number of rounds, or until stable imputed values occur (Raghunathan et al., 2001, p. 87). The normal linear model is as follows, where yi,t is the income at time t:

$$\hat{y}_{i,t} = \hat{\beta}_0 + \hat{\beta}_1 y_{i,1} + \cdots + \hat{\beta}_{t-1} y_{i,t-1} + \cdots + \hat{\beta}_{t+1} y_{i,t+1} + \cdots + \hat{\beta}_7 y_{i,7}$$

The other seven income variables are the independent variables. Note that the imputations had an X that consisted only of a vector of ones.

3.5 Results

Data visualization was used to investigate different properties related to the respondent data, hot deck imputed data, and linear model imputed data. The figures in this section focus on the percentages (point estimates) of the income categories, the standard errors associated with the percentage estimates of the income categories, and the ratio of the MI standard errors divided by the single imputation standard error for the income categories. In addition, the relationship between the point estimates and the standard errors are examined through the 95% confidence intervals for each income category by imputation approach. In all of the presented figures, the imputed estimates include both the imputed values and the nonimputed respondent values.

In all of the figures, estimates are provided by income category level. The NCVS has 14 income category levels; ***Table 3-3*** shows the income range for each level.

Table 3-3. Income categories (Question 12a) in the National Crime Victimization Survey

Household income code	Income level
1	Less than $5,000
2	$5,000 to $7,499
3	$7,500 to $9,999
4	$10,000 to $12,499
5	$12,500 to $14,999
6	$15,000 to $17,499
7	$17,500 to 19,999
8	$20,000 to 24,999
9	$25,000 to $29,999
10	$30,000 to $34,999
11	$35,000 to $39,999
12	$40,000 to $49,999
13	$50,000 to $74,999
14	$75,000 or more

Source: Bureau of Justice Statistics, National Crime Victimization Survey (NCVS), 2010 NCVS-1 basic screen questionnaire.

3.5.1 Consistency of Estimates

To assess the consistency of estimates for each imputation method, the distribution of households by income level was plotted. These estimates included both respondent and imputed values. As a comparison group, the distribution of income for respondents who reported income and did not have their income data imputed was plotted as well. ***Figure 3-2*** shows the results for quarters 1 and 2 of 2010, and ***Figure 3-3*** shows the results for quarters 3 and 4. All of the figures have 13 individual lines each—6 for hot deck (single imputation plus 5 MI data sets), 6 for linear modeling, and 1 for the nonimputed respondents in the comparison group. Some key findings from these two figures include the following:

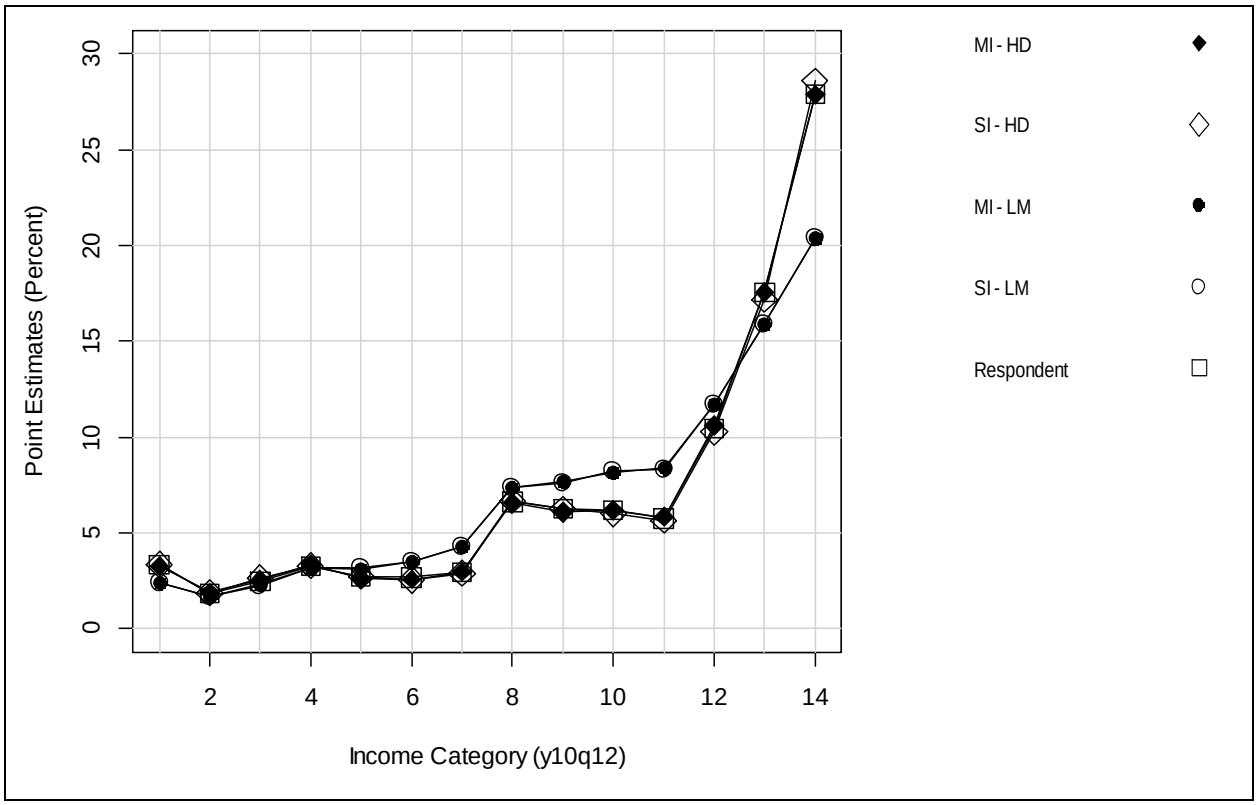
- The multiple imputation approaches produce nearly identical results regardless of the number of imputations conducted. This is why it appears that there is only one line for each of the MI approaches (i.e., the lines are essentially sitting on top of each other).

- The distribution of household income level from the MI-HD approach (lines with diamonds) is nearly identical to the distribution among respondents.

- The MI-LM approach (lines with circles) produces a smoothing effect for the distribution of household income. In other words, missing responses are more likely to be imputed in one of the middle-income categories, reducing the number of observations at both the low and the high ends of the income scale.

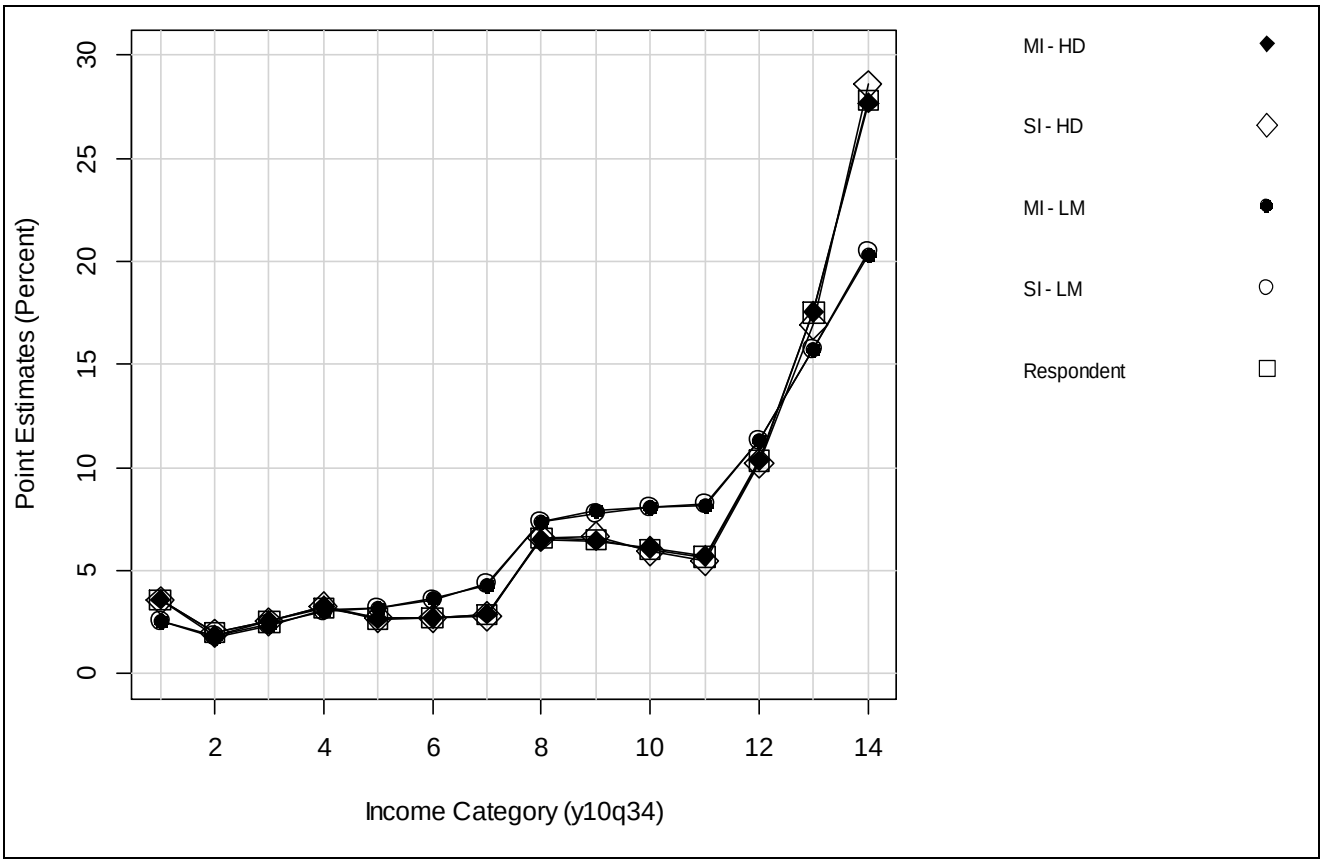
- Both single imputation approaches produce results nearly identical to their multiple imputation counterparts.

Figure 3-2. Percentage point estimates by income category in 2010, quarters 1 and 2



Note: HD, hot deck; LM, linear model; MI, multiple imputation; SI, single imputation.
Source: Bureau of Justice Statistics, National Crime Victimization Survey (NCVS), 2006–2010.

Figure 3-3. percentage point estimates by income category in 2010, quarters 3 and 4



Note: HD, hot deck; LM, linear model; MI, multiple imputation; SI, single imputation.
Source: Bureau of Justice Statistics, National Crime Victimization Survey (NCVS), 2006–2010.

To further illustrate the differences in the estimates between the hot deck and linear model methods, ***Tables 3-4*** and ***3-5*** present the distribution of the nonimputed respondent's household income and the imputed respondent's household income. These tables reinforce that the distribution of respondent's imputed income by hot deck is nearly identical to the distribution of the nonimputed respondent's income, whereas the distribution of respondent's income imputed by the linear model method is not.

Table 3-4. Number and percentage of sample members with reported and imputed income values using single imputation hot deck, 2010

	Number and percentage of National Crime Victimization Survey sample members					
	Non-imputed respondents		Imputed respondents		All respondents	
Income category	Number	Percent	Number	Percent	Number	Percent
Less than $5,000	1,768	3.2	907	3.4	2,675	3.3
$5,000–$7,499	1,006	1.8	509	1.9	1,515	1.8
$7,500–$9,999	1,366	2.5	719	2.7	2,085	2.5
$10,000–$12,499	1,765	3.2	906	3.4	2,671	3.3
$12,500–$14,999	1,471	2.7	738	2.8	2,209	2.7
$15,000–$17499	1,476	2.7	665	2.5	2,141	2.6
$17,500–$19,999	1,593	2.9	721	2.7	2,314	2.8
$20,000–$24,999	3,649	6.6	1,794	6.7	5,443	6.6
$25,000–$29,999	3,522	6.4	1,789	6.7	5,311	6.5
$30,000–$34,999	3,368	6.1	1,505	5.7	4,873	5.9
$35,000–$39,999	3,151	5.7	1,372	5.2	4,523	5.5
$40,000–$49,999	5,762	10.4	2,628	9.9	8,390	10.2
$50,000–$74,999	9,785	17.7	4,301	16.2	14,086	17.2
$75,000 or more	15,682	28.3	8,030	30.2	23,712	28.9

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file..

Table 3-5. Number and percentage of sample members with reported and imputed income values using single imputation linear model, 2010

	Number and percentage of National Crime Victimization Survey sample members					
	Non-imputed respondents		Imputed respondents		All respondents	
Income category	Number	Percent	Number	Percent	Number	Percent
Less than $5,000	1,768	3.2	93	0.3	1,861	2.3
$5,000–$7,499	1,006	1.8	350	1.3	1,356	1.7
$7,500–$9,999	1,366	2.5	508	1.9	1,874	2.3
$10,000–$12,499	1,765	3.2	760	2.9	2,525	3.1
$12,500–$14,999	1,471	2.7	1,091	4.1	2,562	3.1
$15,000–$17499	1,476	2.7	1,432	5.4	2,908	3.5
$17,500–$19,999	1,593	2.9	1,942	7.3	3,535	4.3
$20,000–$24,999	3,649	6.6	2,380	9.0	6,029	7.4
$25,000–$29,999	3,522	6.4	2,791	10.5	6,313	7.7
$30,000–$34,999	3,368	6.1	3,291	12.4	6,659	8.1
$35,000–$39,999	3,151	5.7	3,630	13.7	6,781	8.3
$40,000–$49,999	5,762	10.4	3,711	14.0	9,473	11.6
$50,000–$74,999	9,785	17.7	3,278	12.3	13,063	15.9
$75,000 or more	15,682	28.3	1,327	5.0	17,009	20.8

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010 longitudinal file..

3.5.2 Determining the More Accurate Approach

As shown in ***Figures 3-2*** and ***3-3***, the hot deck and linear model approaches produce different distributions of income. However, because the true value of income for those who had their income data imputed is not known, it cannot be determined conclusively which is the more accurate distribution. Therefore, to determine which approach is more accurate, circumstantial methods must be used. The following are two such methods:

1. Comparison to an external distribution. Under this method, the two imputed distributions are compared to an external estimate of the national household distribution of income. One such source is the ACS, which produces annual estimates of income with a high level of precision (a margin of error of +/- 0.1%) and is, therefore, a good comparison source for the NCVS.

2. A Monte Carlo simulation, in which a fictitious population with known parameters (including income) is created. Then, for a segment of the population (i.e., the

proportion of the population that does not respond to the income question), missing values for the variable of interest (income in this case) are imposed. Next, the missing values are imputed under each of the two approaches. Finally, the bias in the imputed values is computed by comparing the imputed value to the known or actual population estimate. This process is repeated to determine the variance of the bias.

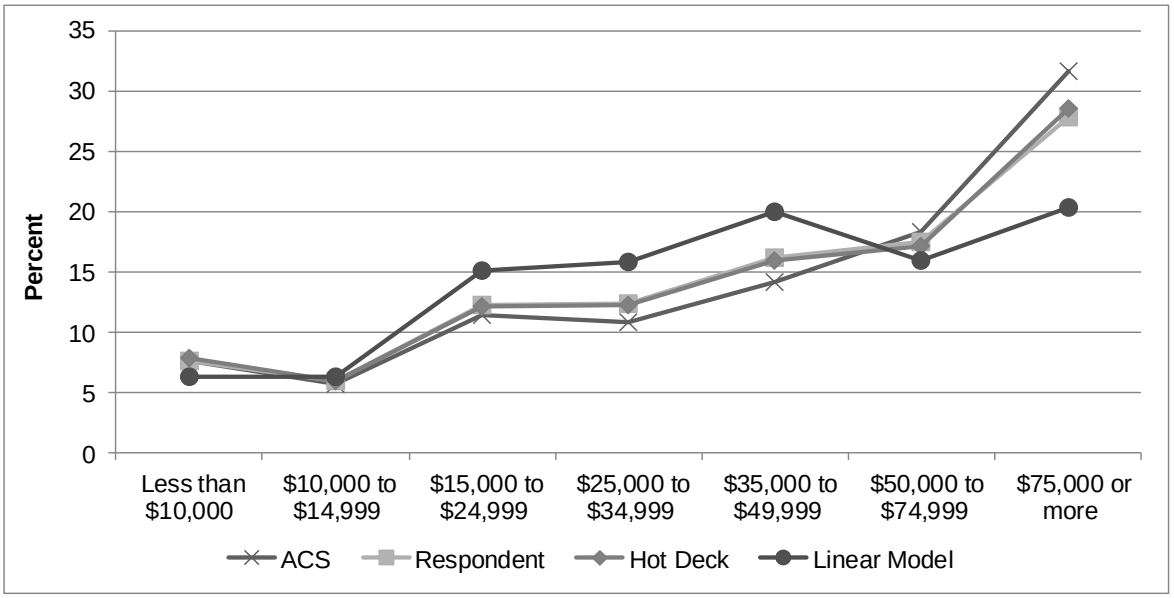
Each of these two methods was explored and the results are described below.

3.5.2.1 Comparison to an External Source

Figure 3-4 shows the results of a comparison between the income distribution estimated by the ACS and three versions of the income distribution estimated using the NCVS:

1. the distribution of respondents only,
2. the distribution of respondents and imputed values using the hot deck approach, and
3. the distribution of respondents and imputed values using the linear model approach.

Figure 3-4. Comparison of income distribution for the 2010 American Community Survey and the Respondent and Imputed Versions of the 2010 National Crime Victimization Survey



Sources: Census Bureau American Community Survey (ACS), 2010; Bureau of Justice Statistics, National Crime Victimization Survey, 2010.

The figure indicates that household income distributions that include data imputed using the hot deck method are more similar to the household income distributions estimated by the ACS than the household income distributions that include data imputed using the linear model approach.

3.5.2.2 Monte Carlo Simulation

Because single and multiple imputation using both the hot deck and the linear model approaches produced different estimates of the income distribution, a Monte Carlo simulation with a known income distribution that is s similar to the NCVS data was used to determine which strategy produced the most accurate distribution. The Monte Carlo simulation used random samples of nonrespondents and different missing data assumptions, including missing completely at random, missing at random, and missing not at random. These random samples were generated from a single population with known income distributions for the time periods in which an NCVS household would have been asked to provide household income. The income variable being imputed is ordinal and has 14 levels corresponding to the 14 levels in the NCVS income variable (see Table 3-3 for definition of categories). The missing values were imputed using each of the two strategies, hot deck and linear model, and using both the single and multiple imputation approaches.

The imputation methods were evaluated using multiple criteria, including a review of the bias, relative bias, confidence interval length, and coverage. The results show that MI-HD had the best coverage. Specifically, the 95% confidence intervals for the MI-HD approach contained the true values more often than the SI-HD, SI-LM, or MI-LM approaches. The SI-HD method performed well in terms of the estimates, bias, and relative bias. The linear models, using both single and multiple imputations, performed poorly in terms of the estimates, as well as in terms of bias, relative bias, confidence intervals, and coverage. More information about the Monte Carlo simulation, including details about the methods and findings, is presented in ***Appendix A.***

3.6 Variability of Imputations

In addition to understanding whether an approach produces accurate estimates, it is important to know the precision, or variability, by which it produces those estimates in order to ensure that, upon repeated application of an approach, the expected estimates will be produced

consistently. In an effort to assess the variability of the imputation procedures being compared, the standard errors were analyzed in two different ways. First, the standard error estimates were plotted by income category for each of the imputation methods. Second, the ratio of the standard error from each of the MI estimates (i.e., the data sets with 5, 10, 15, 20, or 25 imputations) and the standard error from the single imputation estimate were calculated.

Figures 3-5 and ***3-6*** present the standard errors for each imputation procedure by income category for 2010, quarters 1 and 2, and 2010 quarters 3 and 4, respectively. Some key findings from these figures include the following:

- The hot deck imputations had larger standard errors than the linear model imputations.

- There is more variability in the MI-HD imputation than in the MI-LM (as evidenced by the fact that the range of standard errors was larger at any given income category for the MI-HD imputations).

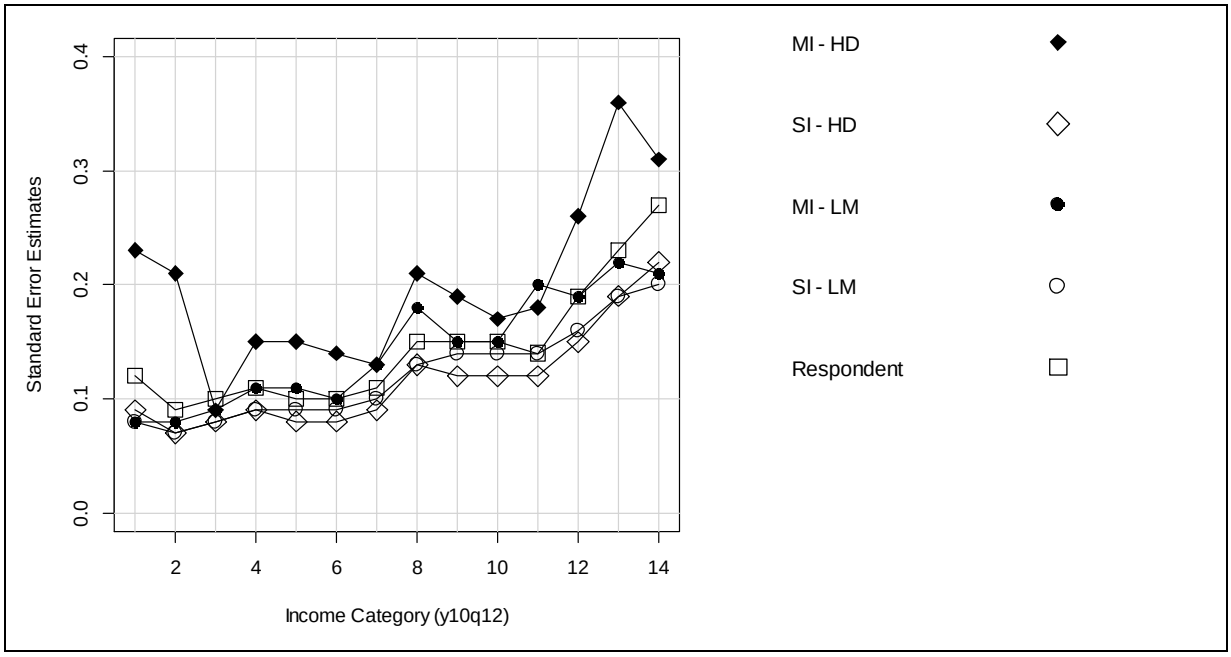
- The number of imputations for each MI approach does not really affect the standard error (as evidenced by the fact that the range of standard errors at each income category is very small).

- The single imputation approaches give standard errors that are too small (i.e., the standard errors are less than the standard errors for the nonimputed respondent, or comparison, estimates). This is because the single imputation approach does not account for the uncertainty in the imputation process. Therefore, the standard error is reduced because of the larger number of income responses (i.e., the denominator of the standard error is larger) but is not consequently increased because of error associated with the imputation process itself.

In Figures 3-5 and 3-6, the respondent standard errors (squares for respondents) fall between the single imputations (diamonds for HD and circles for LM) and multiple imputations (solid diamonds for HD and solid circles for LM). That is, the respondent standard errors are generally higher than the single imputation standard errors and lower than the multiple

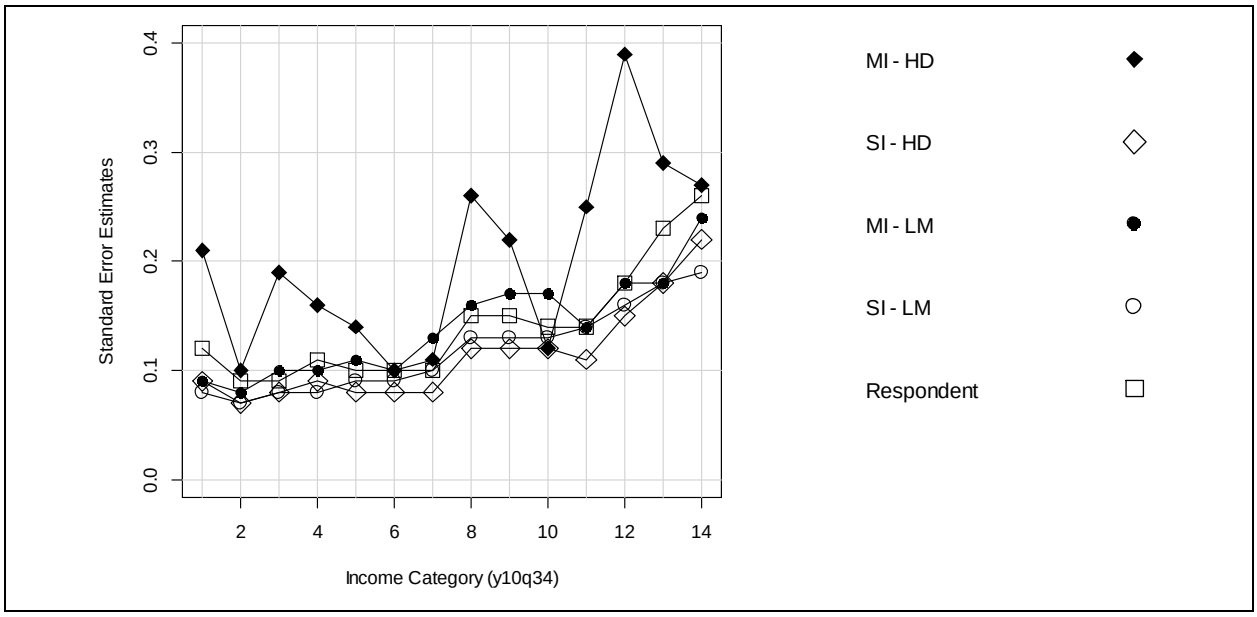
imputation standard errors. This finding makes sense and is consistent with the general impact of imputation on the standard errors of any variable because (1) the multiple imputations account for the uncertainty associated with imputations, so the multiple imputation standard errors would be generally higher than the respondent standard errors, and (2) the single imputations treat the imputed values as respondent values; consequently, the sample size is quite a bit larger than the respondent sample size but does not account for the uncertainty associated with imputation, so the single imputation standard errors would be generally lower than the respondent standard errors. The single imputation approach is often criticized because of this—that is, it makes the standard errors smaller by not accounting for the uncertainty associated with imputing values.

Figure 3-5. Standard error estimates by income category in 2010, quarters 1 and 2



Note: HD, hot deck; LM, linear model; MI, multiple imputation; SI, single imputation.
Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

Figure 3-6. Standard error estimates by income category in 2010, quarters 3 and 4



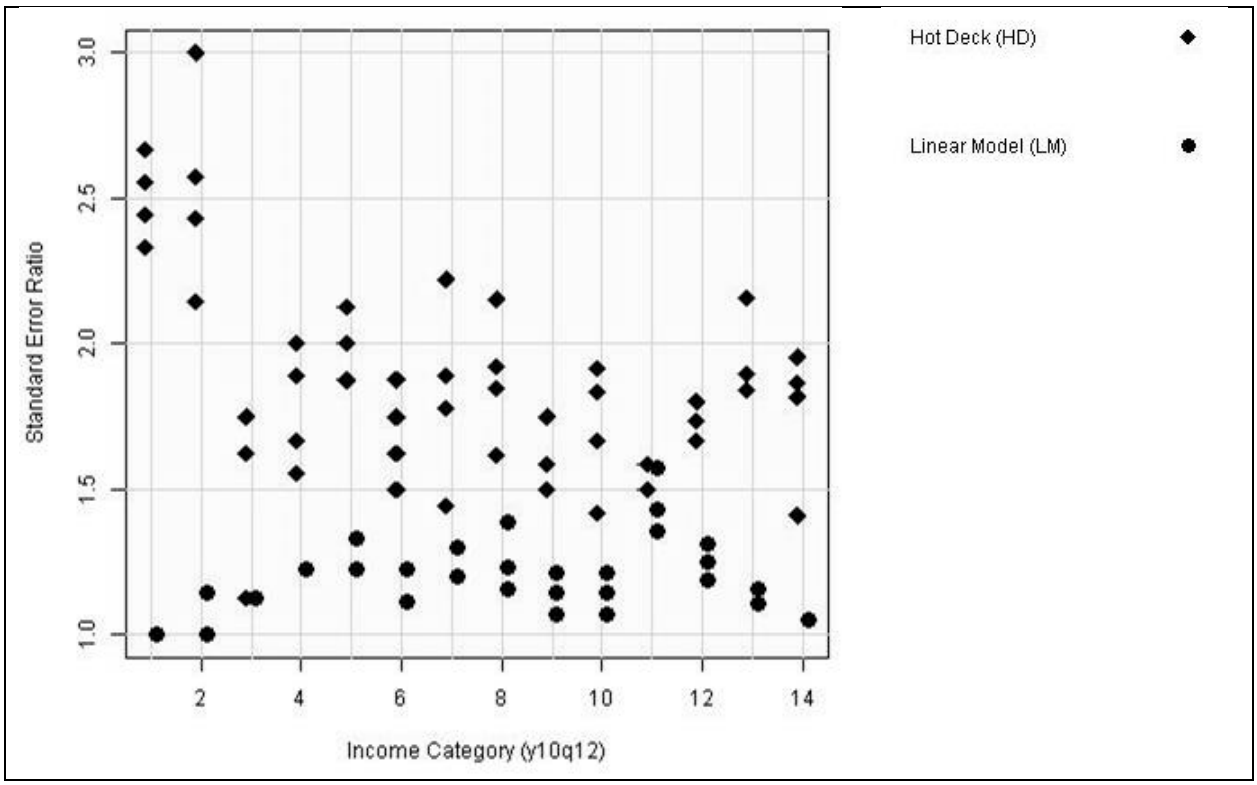
Note: HD, hot deck; LM, linear model; MI, multiple imputation; SI, single imputation.
Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

The above results are further supported by ***Figures 3-7*** and ***3-8***, which present the ratio of the standard errors from each of the MI imputations to the standard errors of the single imputations from the corresponding approaches by income category (i.e., MI-HD to SI-HD and MI-LM to SI-LM) for 2010 quarters 1 and 2 and 2010 quarters 3 and 4, respectively. These two figures show the ratio of the multiple imputation standard error to the single imputation standard error and demonstrate that the single imputation approach underestimates the standard errors. Specifically, in Figures 3-7 and 3-8, each dot in the figures represents one of the MI imputations, where the diamonds are ratios of the hot deck imputations and the black circles represent the ratios of the linear model imputations. There are five dots for each imputation type—one for each number of MI imputations considered (some dots may be on top of each other). This ratio provides good insight because it controls for the relationship between the SI approach and the MI approach and better presents exactly how much larger or smaller the MI standard error estimates are than the single imputation standard error estimates. Some of the key findings from the figures include the following:

- The MI standard errors are always larger than the single imputation standard error imputations (i.e., the ratio values are always greater than one).

- The MI-HD standard errors have more variability than the MI-LM standard errors (as evidenced by the larger ratios for each income category).

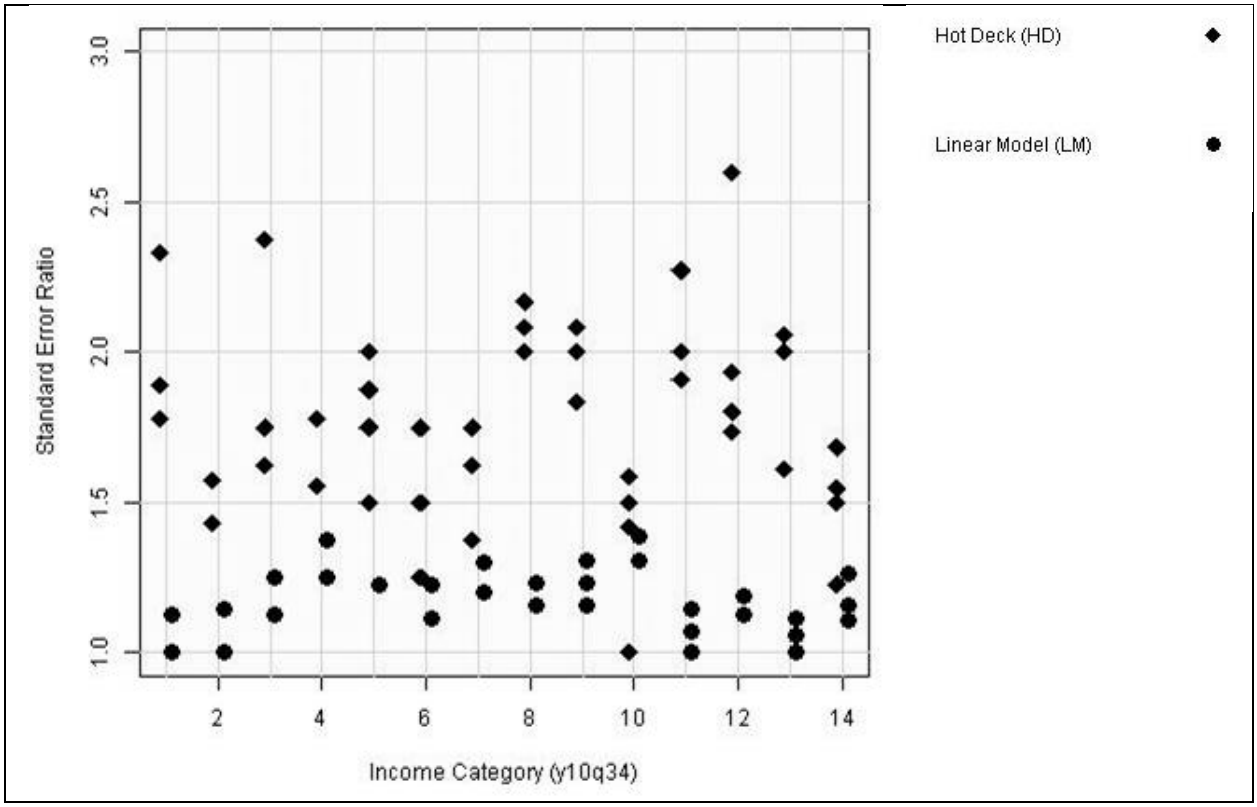
Figure 3-7. Standard error ratio (MI SE/SI SE) by income category in 2010, quarters 1 and 2



Note: For each x-axis line, there are 5 dots for each type of symbol (5 for hot deck and 5 for linear model). If fewer than 5 are visible, it is because some are stacked directly on top of each other.

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

Figure 3-8. Standard error ratio (MI SE/SI SE) by income category in 2010, quarters 3 and 4



Note: For each x-axis line, there are 5 dots for each type of symbol (5 for hot deck and 5 for linear model). If fewer than 5 are visible, it is because some are stacked directly on top of each other.

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

3.7 Relationship Between Point Estimates and Standard Errors

As a final analysis of the two approaches, the relationships between the point estimates and the standard errors were examined. This was done by reviewing the 95% confidence intervals for each income level category by imputation approach. ***Figures 3-9*** and ***3-10*** present these for 2010 quarters 1 and 2 and 2010 quarters 3 and 4, respectively. Within each figure there is a smaller figure for each income category. In addition, each figure has a legend indicating the order in which the 95% confidence intervals for each imputation approach are presented. The lines with diamonds represent the HD imputations, the lines with circles represent the LM imputations, and the lines with the hollow circles represent the respondent (comparison) data. Some key findings from these figures include the following:

- Across most income categories, the 95% confidence intervals for the LM approaches are smaller than for the HD approaches, indicating that the standard errors for the LM approaches are smaller.

- The point estimates (dot in each line) are consistent within a particular approach (HD or LM), indicating that, from a point estimate standpoint, within an imputation approach the number of imputations does not matter.

- The point estimates for the HD approach are very close to the point estimates for the respondent (comparison) sample, whereas the point estimates for the LM approach are different (larger for low-income categories and smaller for high-income categories). In other words, the HD imputations imply that respondents with missing income data have a distribution of income similar to that of the respondents who report income, whereas the LM imputations imply that respondents with missing income data have a distribution of income different from that of the respondents who report income.

Figure 3-9. 95% confidence intervals by income category in 2010, quarters 1 and 2

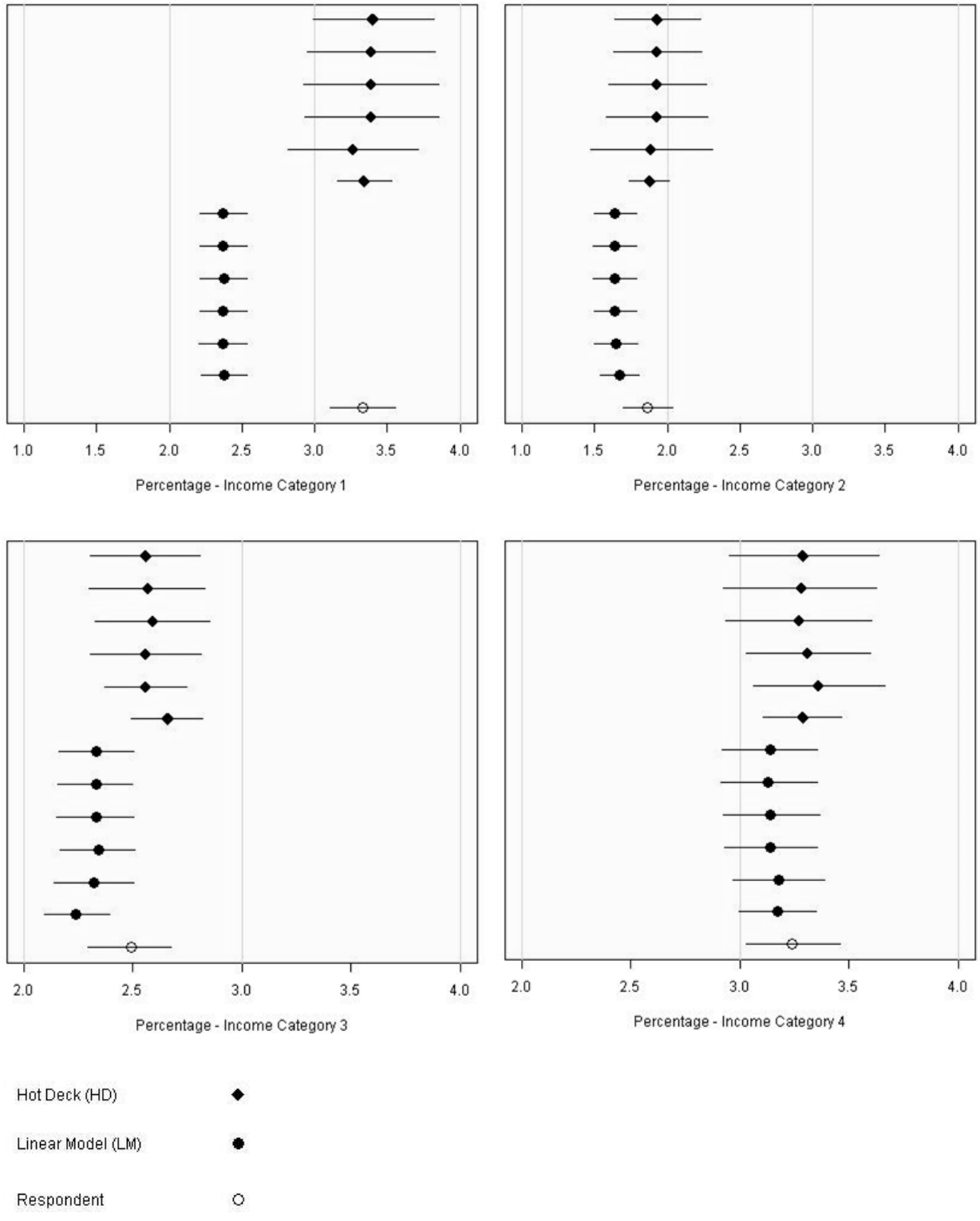
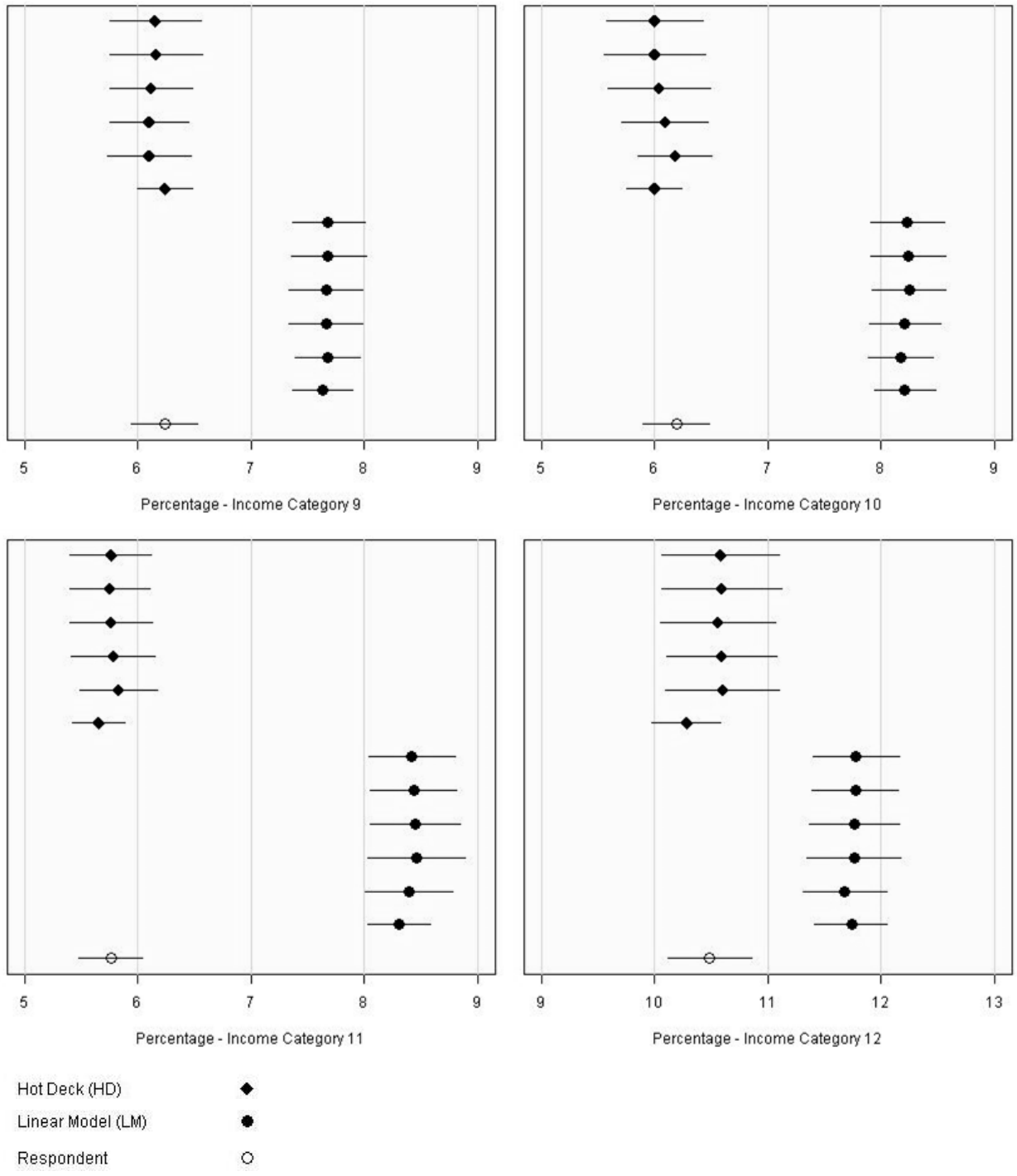


Figure 3-9. 95% confidence intervals by income category in 2010, quarters 1 and 2 (continued)



Note: For each imputation method there are 6 lines representing approaches taken within each method (i.e., single imputation and multiple imputation with 5, 10, 15, 20, and 25 imputations).

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

Figure 3-10. 95% confidence intervals by income category in 2010, quarters 3 and 4

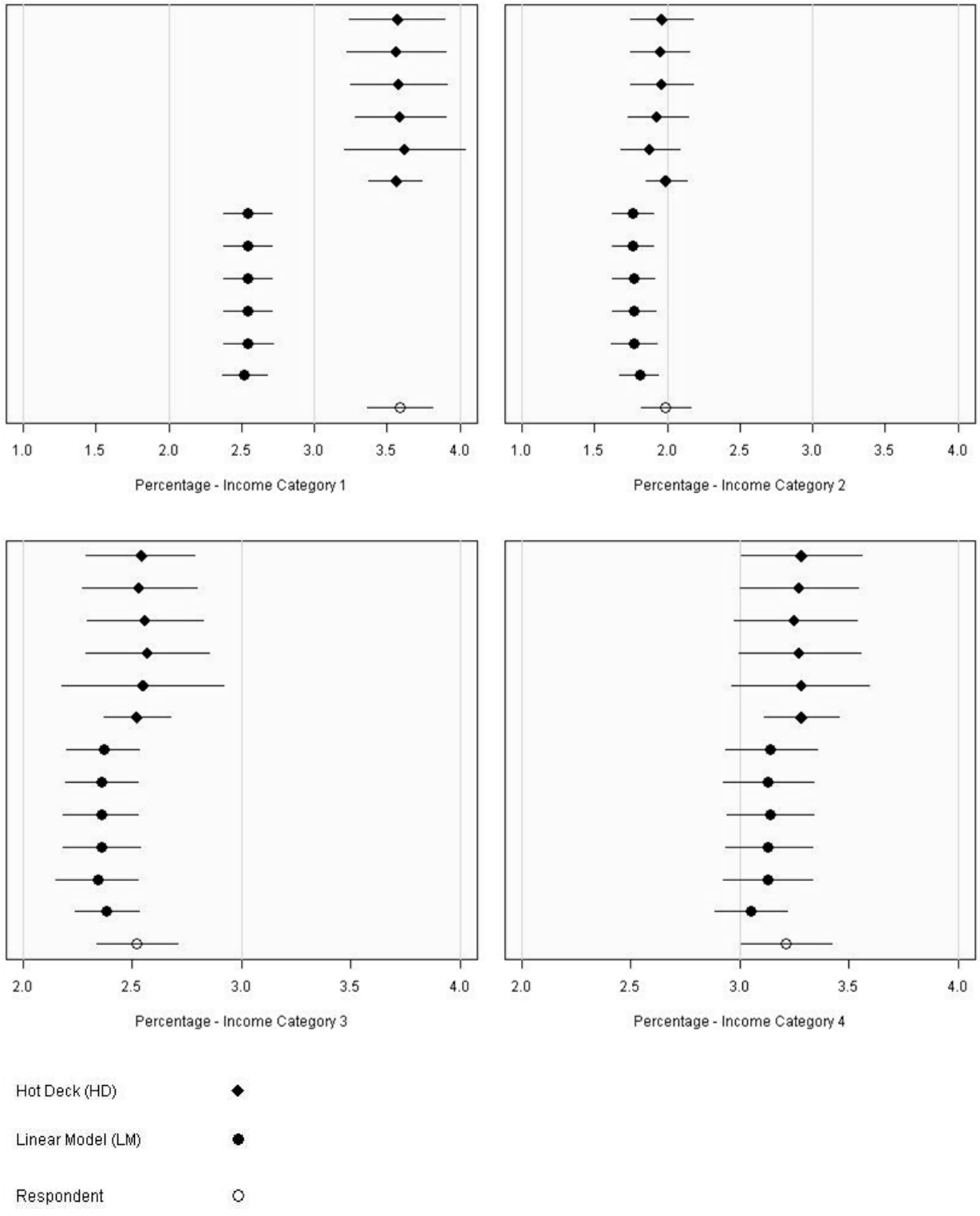


Figure 3-10. 95% confidence intervals by income category in 2010, quarters 3 and 4 (continued)

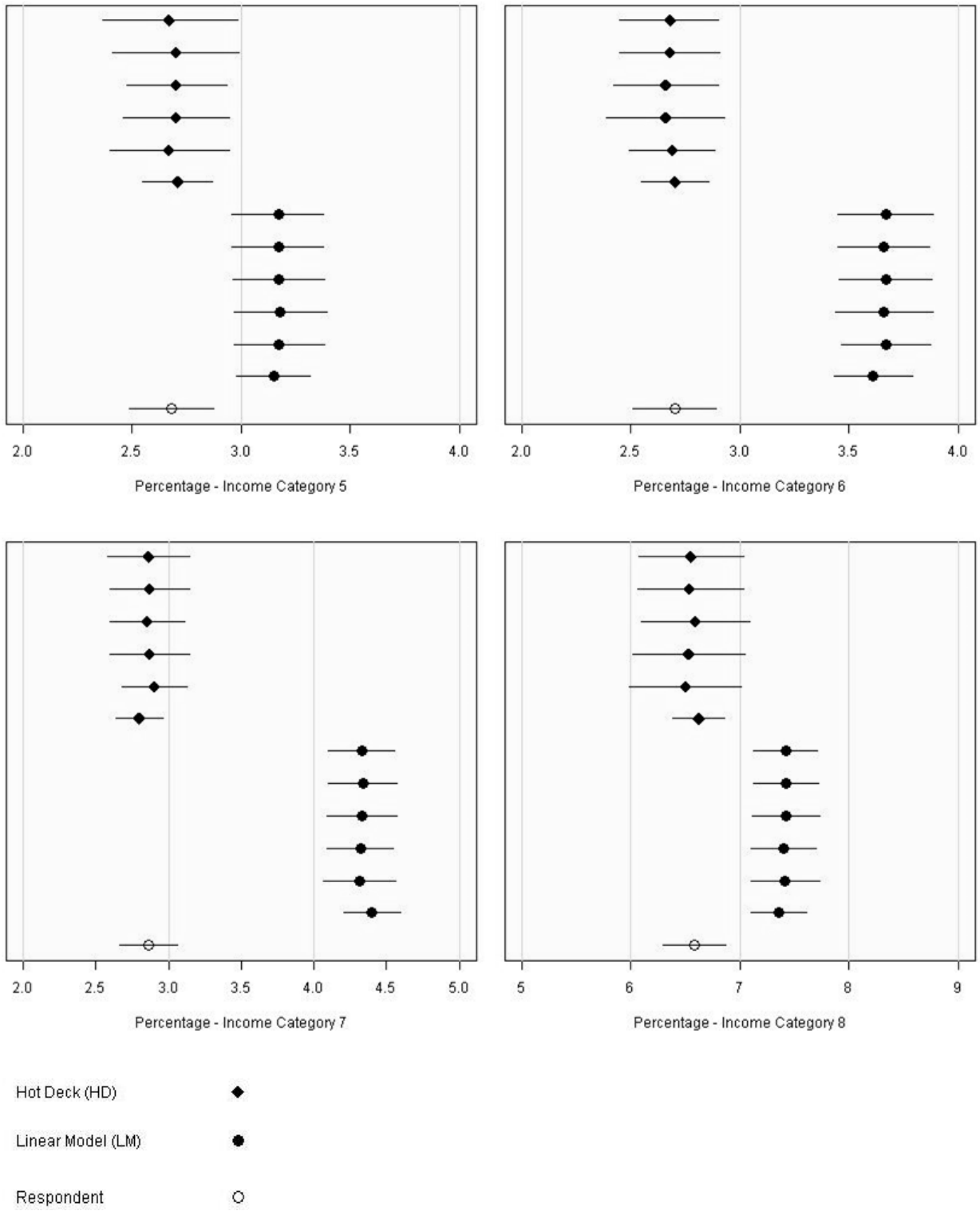


Figure 3-10. 95% confidence intervals by income category in 2010, quarters 3 and 4 (continued)

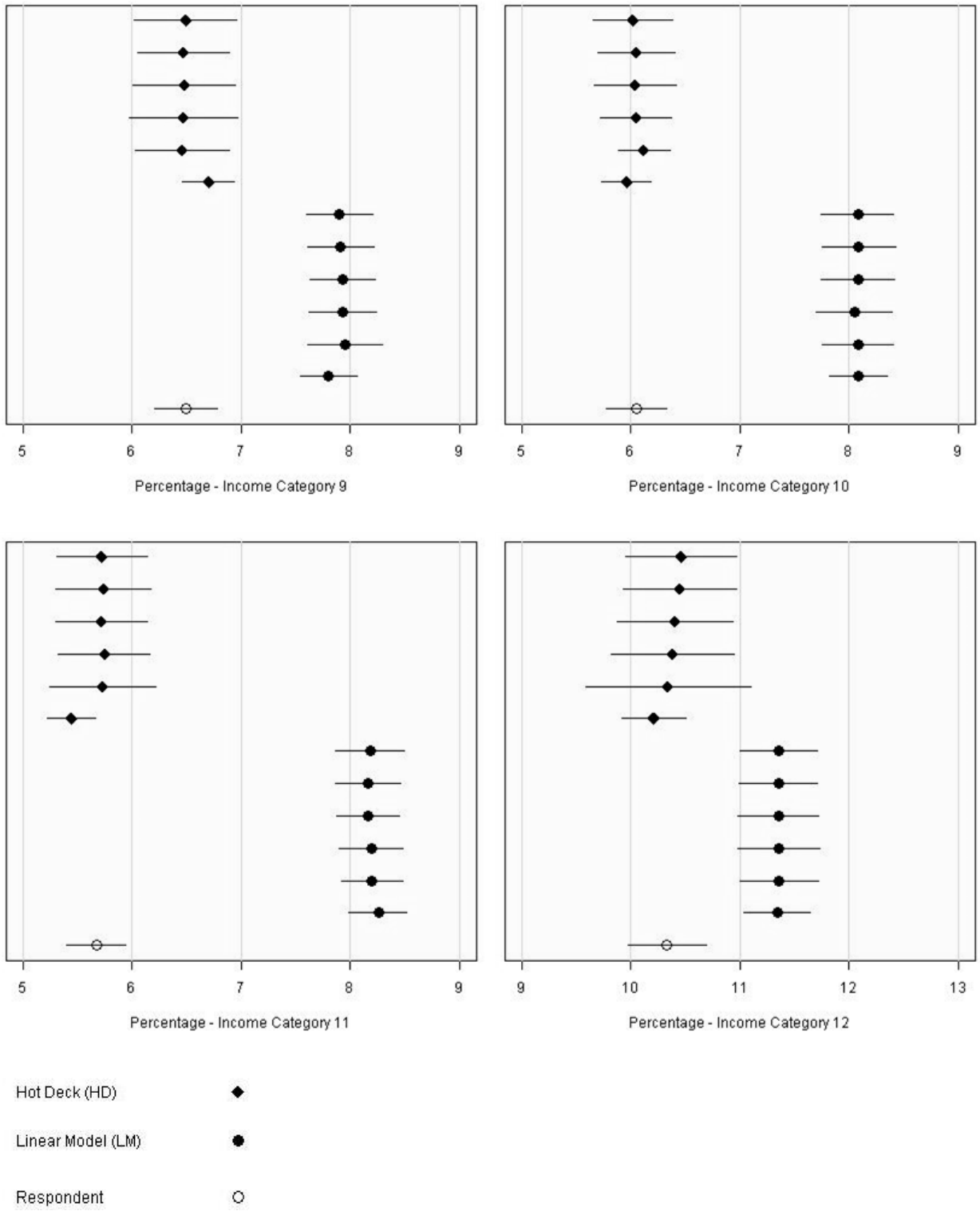
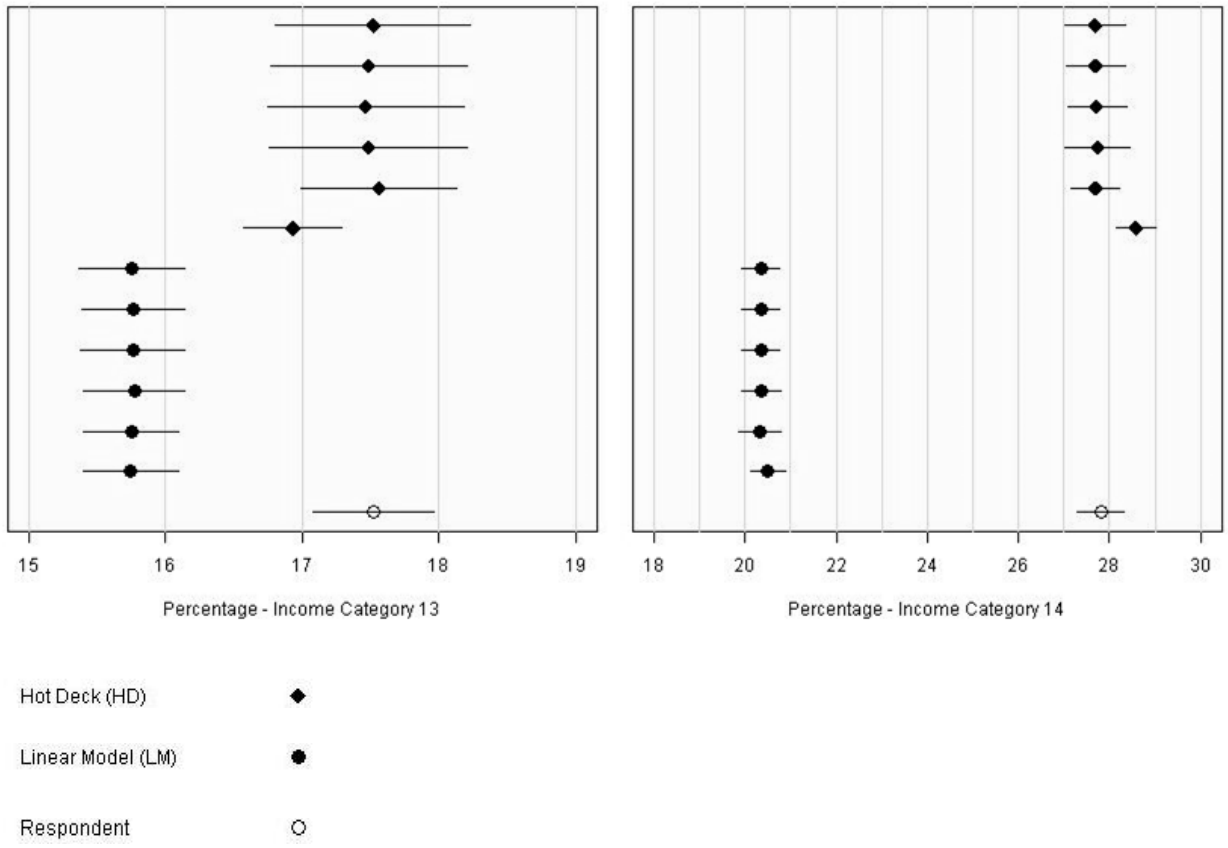


Figure 3-10. 95% confidence intervals by income category in 2010, quarters 3 and 4 (continued)



Note: For each imputation method there are 6 lines representing approaches taken within each method (i.e., single imputation and multiple imputation with 5, 10, 15, 20, and 25 imputations).

Source: Bureau of Justice Statistics, National Crime Victimization Survey, 2006–2010.

3.8 Conclusion

Each of the imputation procedures was rated based on three criteria:

1. Consistency of point estimates
2. Variability of the estimates
3. Ease of implementation

The results of the preceding analyses and the merits of the imputations procedures are summarized below with respect to each of these criteria.

3.8.1 Consistency of Point Estimates

As shown in Figures 3-2, 3-3, 3-9, and 3-10, the point estimates for each income category under the HD approaches tracked very closely with the respondent (comparison) data, whereas the estimates from the LM approaches smoothed out the distribution of income. However, as shown in Figure 3-4, when compared to the ACS, it appears that the hot deck approach better estimates the national distribution of income.

3.8.2 Variability of Estimates

As shown in Figures 3-4 through 3-10, the LM standard errors were consistently smaller across income categories than the standard errors from the HD procedures. Therefore, it is clear that the LM approach is better from a variability standpoint. Furthermore, as shown in Figures 3-5 and 3-6, the standard errors from the single imputation approaches were lower than the standard errors from the respondent (comparison) data alone. The single imputation approach underestimates the standard errors because it does not account for the uncertainty associated with the imputed values. Therefore, a multiple imputation approach provides more plausible standard errors than the single imputation approaches. In addition, as shown in all of the figures, the number of multiple imputations conducted does not greatly alter the point estimates or the variability of the estimates. Therefore, among the MI procedures, there is no need to conduct more than five imputations.

3.8.3 Ease of Implementation and Analysis

As discussed earlier, hot deck imputations are easier to implement than linear model imputations because, unlike the linear model imputations, hot deck imputations do not require special software. Hot deck imputations also have the benefit of always producing an integer value for categorical variables, which removes any potential rounding errors. Furthermore, single imputation approaches are easier to analyze because they do not require special procedures to account for the multiple imputation data sets.

3.9 Recommendations

On the basis of these conclusions, the following recommendations for imputing income data in the NCVS are offered:

- The hot deck approach is recommended. The Monte Carlo simulation clearly demonstrated that the hot deck approach produced an income distribution that more closely mimics the population than does the linear model method. Furthermore, the hot deck distribution more closely followed the distribution from the ACS, indicating that it was closer to the true distribution compared to the linear model method.

- A single imputation procedure is recommended. Although the MI procedures produce more realistic standard errors and had better coverage based on the Monte Carlo simulation (for hot deck), the single imputation approaches are easier to analyze in the NCVS. With three analysis data sets, having to multiply imputed household-level data will be cumbersome analytically and may confuse users given that 1) the general difficulty of handling multiply imputed data, and 2) the incident file would not need to be imputed.

3.10 Validation of Recommendations

Given the recommendation to use the SI-HD method for imputing a respondent's household income category, household income was imputed for three additional survey years: 2008, 2009, and 2011. Earlier years were not imputed because either (1) all interview waves for a household could not be linked using the NCVS public use file because of the Census geography change from 1990 to 2000, or (2) ACS estimates were not available.[4] ***Table 3-6*** compares the distribution of the imputed household income to the distribution of household income reported by the ACS for each of these years. As was the case in Figure 3-4 for survey year 2010, the imputed distributions of household income data for 2008, 2009, and 2011 are similar to the distribution estimated by the ACS. This suggests that if these analyses are repeated for additional years of data, retrospectively or prospectively, the SI-HD will produce accurate results. Significance testing was not performed for this particular analysis for two reasons: (1) the two data sets had large sample sizes and (2) comparing the distributions is of more utility when the substantive differences are relatively small.

[4] The 1-year ACS began in 2005.

Table 3-6. Comparison of the distribution of household income between the National Crime Victimization and American Community Surveys, 2008, 2009, and 2011

	2008		2009		2011	
Income category	NCVS	ACS	NCVS	ACS	NCVS	ACS
Less than $10,000	7.5%	7.2%	7.3%	7.8%	7.8%	7.8%
$10,000–$14,999	5.7	5.4	5.8	5.7	6.1	5.8
$15,000–$24,999	11.9	10.7	11.9	11.2	12.2	11.4
$25,000–$34,999	12.0	10.4	12.7	10.7	12.5	10.6
$35,000–$49,999	16.4	14.2	16.8	14.4	15.7	13.9
$50,000–$74,999	17.4	18.8	17.4	18.3	17.3	18.0
$75,000 or more	29.1	33.4	28.1	31.7	28.5	32.5

Sources: American Community Survey (ACS), 2008, 2009, and 2011; Bureau of Justice Statistics, National Crime Victimization Survey (NCVS), 2008, 2009, and 2011.

REFERENCES

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.* New York, NY: Chapman & Hall.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. In *Proceedings of the Survey Research Methods Section* (pp. 721–726). Alexandria, VA: American Statistical Association.

Engels, J. M., & Diehr, P. (2003). Imputation of missing longitudinal data: A comparison of methods. *Journal of Clinical Epidemiology, 56,* 968–976.

Giusti, C., & Little, R. J. A. (2011). An analysis of nonignorable nonresponse to income in a survey with a rotating panel design. *Journal of Official Statistics, 27,* 211–229.

Home Office. (2011, October). *User guide to Home Office crime statistics*. London, England: Author. Retrieved from http://www.homeoffice.gov.uk/publications/science-research-statistics/research-statistics/crime-research/user-guide-crime-statistics/user-guide-crime-statistics?view=Binary

Iannacchione, V. G. (1982). Weighted sequential hot deck imputation macros. In *Proceedings of the Seventh Annual SUGI—SAS Users Group International Conference* (pp. 759–763). Cary, NC: SAS Institute.

Judkins, D. R. (1997). Imputing for Swiss cheese pattern of missing data. In *Proceedings of the Statistics Canada Symposium 97, New Directions in Surveys and Censuses* (pp. 143–148). Hull, Quebec: Statistics Canada, Methodology Branch.

Little, R., & Su, H. L. (1989). Item nonresponse in panel surveys. In D. Kasprzyk, G. Duncan, G. Kalton, & M. P. Singh (Eds.), *Panel surveys* (pp. 400–425). New York, NY: John Wiley & Sons.

Marker, D., Judkins, D., & Winglee, M. (2002). Large-scale imputation for complex surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge, & R. J. A. Little (Eds.), *Survey nonresponse* (pp. 329–342). New York, NY: Wiley.

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology, 27*(1), 85–95. http://www.isr.umich.edu/src/smp/ive/

Raghunathan, T. E., Solenberger, Peter W., & Van Hoewyk, J. (2002). *IVEware: Imputation and variance estimation software, user guide*. Ann Arbor, MI: University of Michigan, Institute for Social Research, Survey Research Center, Survey Methodology Program.

Research Triangle Institute. (2012). *SUDAAN language manual* (vols. I–II, release 11). Research Triangle Park, NC: Author.

Rubin, D. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley.

Tang, L., Song, J., Belin, T. R., & Unützer, J. (2005). A comparison of imputation methods in a longitudinal randomized clinical trial. *Statistics in Medicine, 24*, 2111–2128.

Twisk, J., & de Vente, W. (2002). Attrition in longitudinal studies: How to deal with missing data. *Journal of Clinical Epidemiology, 55*, 329–337.

U.S. Census Bureau. (2013, March). American Community Survey, 2010 American Community Survey 1-Year Estimates, Table GP03; generated by Marcus Berzofsky; using American FactFinder. Retrieved from http://factfinder2.census.gov

U.S. Census Bureau. (2011). *American FactFinder: Table S1901: Income in the past 12 months (in 2011 inflation-adjusted dollars*). Retrieved from http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_11_1YR_S1901&prodType=table

U.S. Dept. of Justice, Bureau of Justice Statistics. (2013, unpublished). National Crime Victimization Survey, Longitudinal File, 2006-2010. Conducted by U.S. Dept. of Commerce, Bureau of the Census.

Watson, N., & Starick, R. (2011). Evaluation of alternative income imputation methods for a longitudinal survey. *Journal of Official Statistics, 27*, 693–715.

APPENDIX A: INVESTIGATING IMPUTATION APPROACHES FOR SKEWED ORDINAL DATA USING A MONTE CARLO SIMULATION

A.1 Introduction

Each year, the National Crime Victimization Survey (NCVS) data have a high proportion of item nonresponse for the skewed ordinal income variables. In 2010, for example, about a third (32.4%) of weighted respondents did not report household income. Given this high level of nonresponse, an imputation strategy was deemed necessary for correcting item missingness to improve the NCVS's utility and enhance its analytic capacity.

Single and multiple imputation using both the hot deck and the linear model approaches were identified as viable imputation strategies for the NCVS ordinal income variable (see ***Sections 2*** and ***3*** for details on these strategies). However, subsequent analyses using these two methods produced estimates of the income distribution that were different from the actual NCVS data. Therefore, running a Monte Carlo simulation using a known income distribution similar to that of the NCVS data was the next logical step. Broadly, Monte Carlo simulations rely on running multiple replications of a process to solve a problem or generate a distribution of possible outcomes. Monte Carlo methods are mainly used to solve three types of problems: optimization, numerical integration, and generation of samples from a probability distribution. In this analysis, the Monte Carlo simulation was used to generate missing data from a data set with a distribution similar to that of the NCVS sample in an effort to determine which imputation method would yield the most realistic and reliable estimates of the NCVS income distribution.

A.2 Methods

The Monte Carlo simulation was conducted using random samples of item nonrespondents who did not respond to the income question based on different missing data mechanisms that reflect whether the income data are assumed to be (1) missing completely at random (MCAR); (2) missing at random (MAR); and (3) missing not at random (MNAR). The random samples were generated from a single population with known income distributions for the time periods in which an NCVS household would have been asked to provide household income. The NCVS asks respondents to provide household income information every other interview. Therefore, respondents could provide household income once (for those who entered

in the fourth wave), twice (for those who entered in the third wave), three times (for those who entered in the second wave), or four times (for those in the first wave). The imputation methods used were the sequential regression multivariate imputation, or the linear modeling approach (Raghunathan et al., 2001, 2002), and the weighted sequential hot deck (WSHD; Cox, 1980; Research Triangle Institute, 2012). For both the linear model and hot deck approaches, single and multiple imputations were implemented. Standard estimation methods were used for both the single and multiple imputation estimators (Rubin, 1987; Research Triangle Institute, 2012). The missing values were imputed under four strategies:

1. single imputation linear model (SI-LM)
2. multiple imputation linear model (MI-LM)
3. single imputation hot deck (SI-HD)
4. multiple imputation hot deck (MI-HD)

The imputation methods were evaluated across multiple criteria, including bias, relative bias, confidence interval length, and coverage. Of these four criteria, coverage was the primary evaluation method because it measures the proportion of confidence intervals that contain the true value. For 95% confidence intervals, it is expected that the proportion of confidence intervals containing the true value would be about 95%; thus, the coverage properties of the different imputation methods should be 95% or higher. A coverage value for a particular imputation method that is higher than 95% suggests that method to be a methodologically stronger approach for imputing the data.

A.2.1 Imputation Methods

A.2.1.1 Weighted Sequential Hot Deck

The WSHD imputation method (Cox, 1980; Research Triangle Institute, 2012) consisted of initial and cycling phases of imputation. In the initial phase, the starting point of the imputation process was a rectangular data set of 35,000 rows representing households and four columns representing the ordinal income variables for the four time periods. For any household that had at least one valid ordinal income variable value and other missing values, household

imputation was implemented. Household imputation (i.e., row imputation) is implemented using the following steps:

1. The mean of the observed ordinal income values for the household is calculated,
2. the mean is rounded to the nearest integer, and
3. the rounded mean value is filled in for any missing values for the household.

After the household imputation, any remaining missing values were imputed using time period imputation. Time period imputation (i.e., column imputation) is implemented using the following steps:

1. An income value is selected randomly from the distribution of observed ordinal income values for a particular time period, and
2. the randomly selected income value is assigned to a household with a missing income value for the time period.

After the time period imputation, there were no missing values in the rectangular data set. For the single imputation, there was one data set. For the multiple imputation, a specified number of data sets were created. For multiple imputation, 5, 10, 15, and 20 data sets were created. The multiple imputation data sets were created from the original data set by taking a with-replacement sample that had the same number of families as the original data set. The imputation was then applied to these data sets independently.

For each data set created, whether a single data set or multiple imputation data set, the cycling phase was implemented. The term cycling refers to the cyclic n-partition hot deck (Marker, Judkins, & Winglee, 2002)[5] and involves iteratively cycling through n-partition hot decks. This approach is generally based on Bayesian methods and has semiparametric features of the hot deck method (Marker, Judkins, & Winglee, 2002: 334). Thus, there are no strong assumptions made about distribution shapes or about prior distributions for parameters, and

[5] David Judkins is currently (2012 Joint Statistical Meetings) referring to this as *p*-cyclic partition hot deck. He changed from *n* to *p* because *n* is often used to denote the number of observations and *p* the number of variables.

careful choices are made about which features of the covariance structure deserve to be preserved (Marker, Judkins, & Winglee, 2002, p. 334). This cyclic approach is similar to the University of Michigan's IVEware software (Raghunathan et al., 2001) iterative procedure and can be done with or without MI.

The WSHD methodology was implemented during the cycling phase. The WSHD methodology replaces missing data with valid data from a donor record within an imputation class. It incorporates sorting on specified variables within an imputation class for additional control and uses the value of an appropriate weight variable of each record, which includes both donors and recipients, in the donor selection process.

For this study, the imputation classes were formed by cross-classifying the two ordinal income variables closest to the ordinal income variable to be imputed. For example, for the first time period ordinal income variable, the cross-classified variables were the second and third time period ordinal income variables. This convention carried forward through the second, third, and fourth time period ordinal income variables.

In this simulation, no sorting variables were used. Within each imputation class, the hot deck process was sequential because the order of the recipients and donors in the file was important. Weights were used to determine the relative sizes of donors and recipients. The weights were set to one for both donors and recipients. Recipient weights were scaled to the same overall size of the donor weights, and the recipient and donor weights were aligned to be parallel with each other. The recipient weights divided the donor weights into zones for each recipient based on recipient weight size. The possible donors for the recipient were the donors that had a portion or all of their weight in the zone created by the recipient weight. A donor was randomly selected based on the relative sizes of the donor weights in the zone for the recipient.

A.2.1.2 Sequential Regression Multivariate Imputation

The linear model approach used the sequential regression multivariate imputation method (Raghunathan et al., 2001) implemented in IVEware (Raghunathan et al., 2002). This method drew imputed values for the ordinal income variable, Y_j, at round (t+1) from a predictive

distribution corresponding to conditional density. The conditional densities for the ordinal income variable for each of the four time periods, g_j, are presented below.

$$g_1\left(Y_1|Y_2^{(t)}, Y_3^{(t)}, Y_4^{(t)}, \boldsymbol{\varphi}_1\right),$$

$$g_2\left(Y_2|Y_1^{(t+1)}, Y_3^{(t)}, Y_4^{(t)}, \boldsymbol{\varphi}_2\right),$$

$$g_3\left(Y_3|Y_1^{(t+1)}, Y_2^{(t)+1}, Y_4^{(t)}, \boldsymbol{\varphi}_3\right), \text{ and}$$

$$g_4\left(Y_4|Y_1^{(t+1)}, Y_2^{(t+1)}, Y_3^{(t+1)}, \boldsymbol{\varphi}_4\right).$$

The conditional density, $g_{j,}$was specified by the normal linear regression model, and $\boldsymbol{\varphi}_j$ was the vector of unknown regression parameters with a wide range of possible values. Specifically, the new imputed values for a variable were conditional on the previously imputed values of other variables and the newly imputed values of variables that preceded the currently imputed variable (Raghunathan, et al., 2001, p. 88). Because IVEware does not have a cumulative logistic regression model, the normal linear regression model that was restricted to impute values between 0.5 and 14.5 was used. Once the values were imputed, the values were rounded to the nearest integer.

A.2.2 Estimation Methods

Standard estimation techniques were used to calculate estimates for the variables that were imputed using single imputation. Multiple imputation estimators (Research Triangle Institute, 2012; Rubin, 1987) were used to calculate estimates for variables imputed using multiple imputation. In the following formulas, *m* was the number of multiple imputations, θ was the parameter of interest, $\hat{\theta}_i$ was the estimate of the parameter of interest from the *i*th data set, and $\hat{V}\left(\hat{\theta}_i\right)$ was the variance estimate for the estimate $\hat{\theta}_i$ from the *i*th data set. The multiple imputation estimator for the point estimate was the mean of the point estimates from each of the *m* imputations. Specifically, the multiple imputation point estimator, $\hat{\theta}_M$, was

$$\hat{\theta}_M = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i.$$

Finally, the multiple imputation variance estimator for the multiple imputation estimate $\hat{\theta}_M$ was the mean of the variance estimates from each of the *m* imputations and the number of

imputations plus one divided by the number of imputations times the variance of the *m* point estimates. Specifically, the multiple imputation variance estimator for the multiple imputation point estimate, $\hat{V}_M(\hat{\theta}_M)$, was

$$\hat{V}_M(\hat{\theta}_M) = \frac{1}{m}\sum_{i=1}^{m}\hat{V}(\hat{\theta}_i) + \frac{m+1}{m}\left\{\frac{1}{m-1}\sum_{i=1}^{m}(\hat{\theta}_i - \hat{\theta}_M)^2\right\}.$$

A.2.3 Evaluation Methods

The bias, relative bias, confidence interval length, and coverage were analyzed as a means of evaluating each imputation method. These criteria are described in more detail below. As noted earlier, the primary evaluation metric was coverage.

A.2.3.1 Bias

The bias was calculated as the average of the differences between the estimated values and the true value. That is, the bias, b, was calculated as

$$b = \frac{1}{m}\sum_{i=1}^{m}\hat{\theta}_i - \theta,$$

where *m* was the number of iterations of the Monte Carlo simulation, $\hat{\theta}_i$ was the estimated value for the *i*th iteration, and θ was the true value.

A.2.3.2 Relative Bias

The relative bias was calculated as the average of the differences between the estimated values and the true value, divided by the true value. The relative bias, *rb*, was calculated as

$$rb = \frac{1}{m}\sum_{i=1}^{m}\frac{\hat{\theta}_i - \theta}{\theta},$$

where *m* was the number of iterations of the Monte Carlo simulation, $\hat{\theta}_i$ was the estimated value for the *i*th iteration, and θ was the true value.

A.2.3.3 Confidence Interval Length

For a 95% confidence interval, the confidence interval length was calculated as the average of the standard errors multiplied by 3.92. The confidence interval length, *cil*, was

$$cil = \frac{1}{m}\sum_{i=1}^{m} 3.92 \cdot se_i,$$

where *m* was the number of iterations of the Monte Carlo simulation and se_i was the standard error associated with the *i*th estimate.

A.2.3.4 Coverage

For a 95% confidence interval, the coverage was the proportion of times that the confidence interval included the true value. The coverage, *cov*, was

$$cov = \frac{1}{m}\sum_{i=1}^{m} I_{[\hat{\theta}_i - 1.96se_i, \hat{\theta}_i + 1.96se_i]}(\theta),$$

where $I_{[\hat{\theta}_i - 1.96se_i, \hat{\theta}_i + 1.96se_i]}(\theta)$ is the indicator value denoting whether or not the confidence interval includes the true value θ. The indicator variable, $I_{[\hat{\theta}_i - 1.96se_i, \hat{\theta}_i + 1.96se_i]}(\theta)$, is

$$I_{[\hat{\theta}_i - 1.96se_i, \hat{\theta}_i + 1.96se_i]}(\theta) = \begin{cases} 1, if \hat{\theta}_i - 1.96se_i \leq \theta \leq \hat{\theta}_i + 1.96se_i \\ 0, otherwise \end{cases},$$

where $\hat{\theta}_i$ was the estimated value for the *i*th iteration, θ was the true value, and se_i was the standard error associated with the *i*th estimate.

A.3 Data

The analysis used generated data. The data set from which the true value was calculated is referred to as the "known" data set, which means that it was the complete data with known parameters of interest before any missing values were introduced. The known data set was created to represent a rotating panel design with four time periods, which represent the points at which income is asked across the NCVS data collection waves. Once the known data set was created, three types of missing data mechanisms (i.e., MCAR, MAR, and MNAR) were implemented.

A.3.1 Known Data

The known data set had 35,000 observations, each with a unique numbered identifier. For each observation, four time variables were created. The time variables were standard normal random variables correlated over the four time periods. Time 1 and Time 2, Time 2 and Time 3, and

Time 3 and Time 4 each had a theoretical correlation coefficient of 0.95. ***Table A.3-1*** shows the empirical correlation coefficients from the known data.

Table A.3-1. Empirical correlation coefficients from the known data

	Time 1	**Time 2**	**Time 3**	**Time 4**
Time 1	1.00	0.95	0.90	0.86
Time 2		1.00	0.95	0.90
Time 3			1.00	0.95
Time 4				1.00

From each of the standard normal time variables, an ordinal income variable with the 14 NCVS income categories was created. Each of the time variables was independently sorted in ascending order. Once sorted, the income variable values were assigned sequentially. The size of the income categories was based on actual income categories observed from the NCVS data set. ***Table A.3-2*** shows the ordinal income distribution for each ordinal income variable created from the standard normal time variable.

Table A.3-2. Ordinal income distribution created from a standard normal variable

Income code	Number of observations	Cumulative number of observations	Percentage of the data
1	350	350	1
2	350	700	1
3	350	1050	1
4	700	1750	2
5	700	2450	2
6	1,050	3500	3
7	1,050	4550	3
8	1,400	5950	4
9	1,750	7700	5
10	2,100	9800	6
11	2,800	12,600	8
12	3,500	16,100	10
13	7,350	23,450	21
14	11,550	35,000	33

After the ordinal time variables were created, the last step for the known data was to create the panel membership groups for the rotating panel. Four panel membership groups were created to correspond to the four ordinal income variables. Each panel membership group contained 20,000 observations. ***Table A.3-3*** shows the pattern and size of the panel membership groups.

Table A.3-3. Panel membership group

Group 1	Group 2	Group 3	Group 4	Number of observations
0	0	0	1	5,000
0	0	1	1	5,000
0	1	1	1	5,000
1	0	0	0	5,000
1	1	0	0	5,000
1	1	1	0	5,000
1	1	1	1	5,000

A.3.2 Missing Data

For each missing data mechanism (i.e., MCAR, MAR, and MNAR), the level of missingness was targeted to be 0.33 or 33%, because this was the approximate proportion of income missingness in the 2010 NCVS data. One hundred iterations were run for each type of missing data.

A.3.2.1 Data Missing Completely at Random

Data MCAR means the missingness does not depend on any other variables in the data set or the variable to be imputed. Bernoulli random variables are commonly used to identify which original values of the ordinal income variable values should be set to missing and which should be retained as respondent values for the Monte Carlo simulations. Specifically, a Bernoulli random variable has two possible outcomes: success or failure. In the analysis, when the Bernoulli random variable was a success the value was coded as one and the ordinal income variable was set to missing. When it was a failure, it had a value of zero and the ordinal income variable was retained. The probability mass function for a Bernoulli random variable, *p(i)*, was

$$p(i) = P[X = i] = \begin{cases} p, X = 1 \\ 1 - p, X = 0 \end{cases},$$

where p is the probability of success and $1-p$ is the probability of failure. For the Monte Carlo simulation, success was defined as a missing value. That is, for each of the specific time periods, a Bernoulli random variable with a probability of success equal to 0.33 was generated—that is, the probability of a missing value was 0.33. If the value of the Bernoulli random variable was one, the income value was set to missing. Otherwise, the income value was retained.

A.3.2.2 Data Missing at Random

Data MAR means that the missingness depends on other variables in the data set but does not depend on the variable to be imputed. To create the MAR data, the closest ordinal income variable to the variable to be imputed was used. Generally, the value used in this process was the income reported in the previous time period. Next, the lowest three (i.e., income levels 1, 2, and 3) and highest three ordinal income categories (i.e., income levels 12, 13, and 14) for the closest ordinal income variable were set to have a higher level of missingness for the ordinal income

variable to be imputed. As noted in ***Section A.2***, the specific levels of missingness for the income categories were set so the overall level of missingness was 0.33. In general, the higher-level-income sample members were less likely to respond to the income question. This was also true for lower-income sample members but to a lesser extent.

To create the missing values for the first time period, the income categories of the second time period when data were MAR were used. To meet the conditions of MAR and an overall level of missingness of 0.33, the levels of missingness for the two groups of income categories have to be different and the overall level of missingness between them must average the level of missingness (i.e., 0.33). Thus, across the income categories 1, 2, 3, 12, 13, and 14 in the second time period, a Bernoulli random variable with a probability of success equal to 0.39 was generated for the first time period. From the other income categories in the second time period (4–11), a Bernoulli random variable with probability of success equal to 0.27 was generated for the first time period. If the value of the Bernoulli random variable was one, the corresponding income value in the first time period was set to missing. Otherwise, the income value from the first time period was retained. For time periods 2, 3, and 4, the same process was followed.

A.3.2.3 Data Missing Not at Random

Data MNAR means that the missingness may or may not depend on other variables within the data set but is dependent on the variable to be imputed. For this study, to create the MNAR data, the missingness was directly related to the value of the ordinal income variable to be imputed. Generally, higher-level-income sample members are less likely to respond to the income question. Therefore, the level of missingness increases with the amount of income—that is, the higher the value of the ordinal income variable, the more missingness. Consistent with the other missing data mechanisms, specific levels of missingness for the income categories were set so that the overall level of missingness was 0.33.

For the specific time period for data that were not MAR, based on the value of the ordinal income variable, a Bernoulli random variable was created with probability of success equal to the value of the ordinal income variable multiplied by 0.03. If the value of the Bernoulli random variable was one, the income value was set to missing. Otherwise, the income value was

retained. ***Table A.3-4*** shows the value of the ordinal income variable and the proportion missing based on the value of the ordinal income variable.

Table A.3-4. Ordinal income variable and proportion missing

Ordinal income variable	Proportion missing
1	0.03
2	0.06
3	0.09
4	0.12
5	0.15
6	0.18
7	0.21
8	0.24
9	0.27
10	0.30
11	0.33
12	0.36
13	0.39
14	0.42

A.4 Results

Each aspect of the data investigated (i.e., bias, relative bias, confidence interval length, and coverage) had three groups, each consisting of four smaller figures. The three groups were based on the missing data mechanism (i.e., MCAR, MAR, and MNAR). Within each group, a corresponding figure was created for each specific time period (1–4), but for the purposes of brevity, not all of the figures are presented. The true values were only represented in the estimates figure.

A.4.1 Estimates

Figure A.4-1 shows the percentage point estimates for the categories of the ordinal income variable. For all three missing data mechanisms (i.e., MCAR, MAR, and MNAR), for all four time periods for each missing data mechanism, the pattern of estimates are virtually the same. Figure A.4-1 shows the MNAR case for the first time period, and variability among the estimates is evident. There is "over-plotting" within this figure. Specifically:

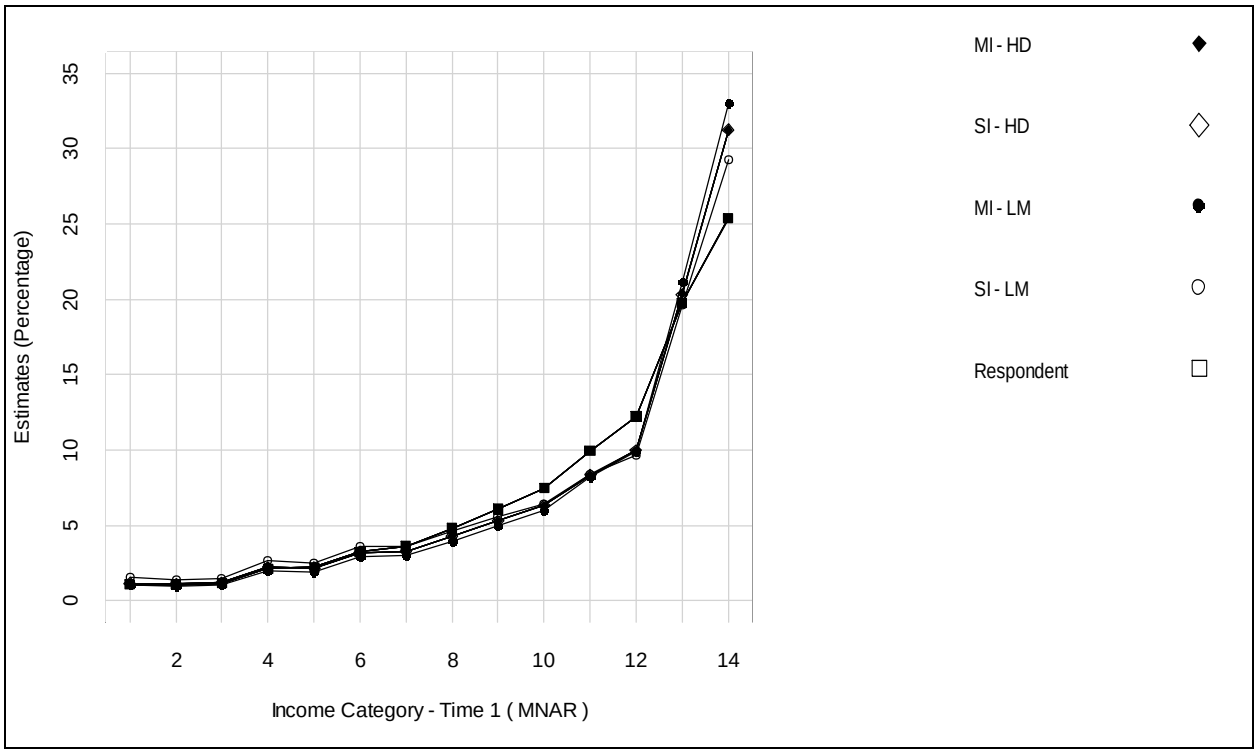
- The MI-LM estimates are nearly the same as the SI-LM estimates and differ from the other estimates.

- The linear model estimates are slightly higher than the true values for ordinal income categories 7–12 and slightly lower than the true values for ordinal income categories 13 and 14.

- The respondent, SI-HD, and MI-HD estimates are similar to the true values.

Although not shown, in both the MCAR and MAR figures, there was less variability in the estimates than is depicted for MNAR in Figure A.4-1. In short, the lines for the different estimates followed the true values more closely in the MCAR and MAR figures. In the case of the MAR, there was less deviation between the true values and the linear model estimates. For the MCAR case, the linear model estimates were closer to the true values with some deviation.

Figure A.4-1. Ordinal income category estimates under missing not at random (MNAR) missing data mechanism



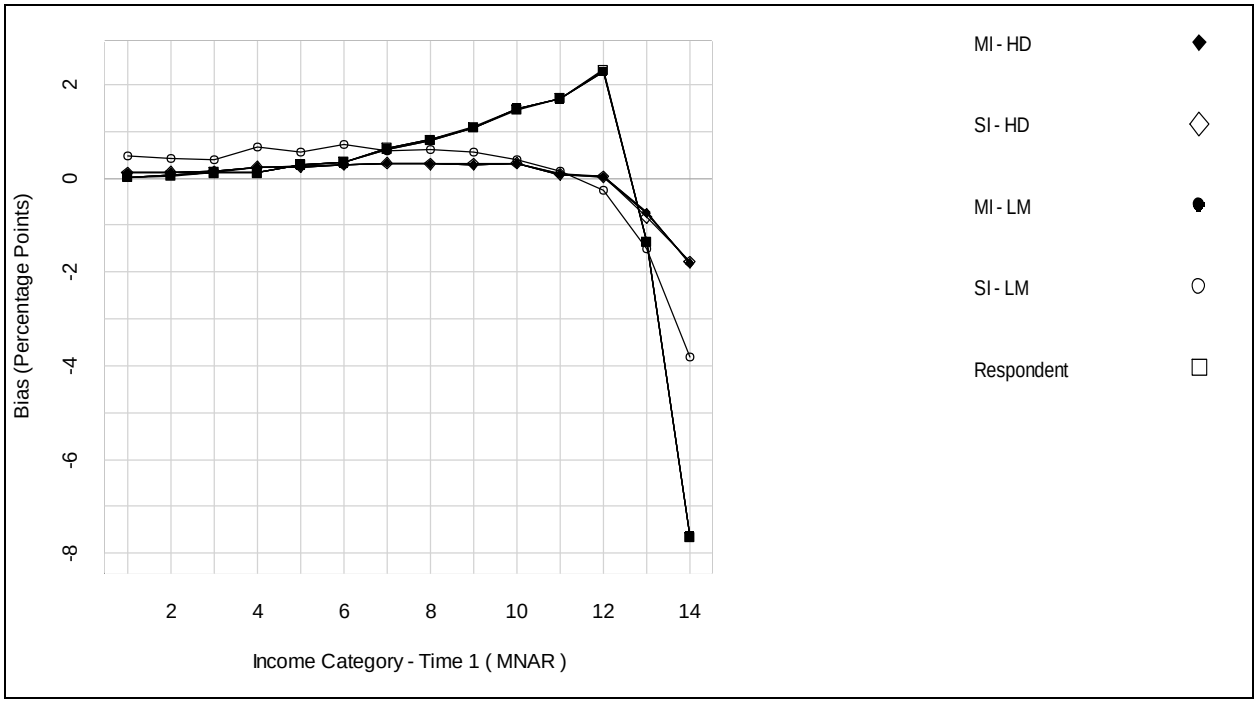
A.4.2 Bias

Bias is the difference between the estimate and the true value. ***Figure A.4-2*** shows the bias in the ordinal income category estimates. The pattern of estimates was similar across all three missing data mechanisms and all four time periods. Figure A.4-2 presents the MNAR case for the first time period, and the following observations may be made:

- There is evident bias for both the SI-LM and the MI-LM.
- Generally, the positive bias starts at ordinal income category 7, increases as the ordinal income categories increase up to ordinal income category 12, and decreases from ordinal income category 12 to 14.
- Ordinal income categories 13 and 14 have considerable negative bias.
- For the SI-HD and MI-HD, there is negative bias for ordinal income categories 13 and 14.
- The respondent bias is more negative than the hot deck bias levels but not as negative as the linear model bias levels.

Although not depicted, the bias for the MAR figure had a pattern similar to that of the MNAR bias, although the negative bias was not as severe. In the MCAR figure, only the SI-LM and MI-LM showed any bias, and it was smaller than the MNAR and MAR bias.

Figure A.4-2. Ordinal income variable category bias under missing not at random (MNAR) missing data mechanism for the first time period

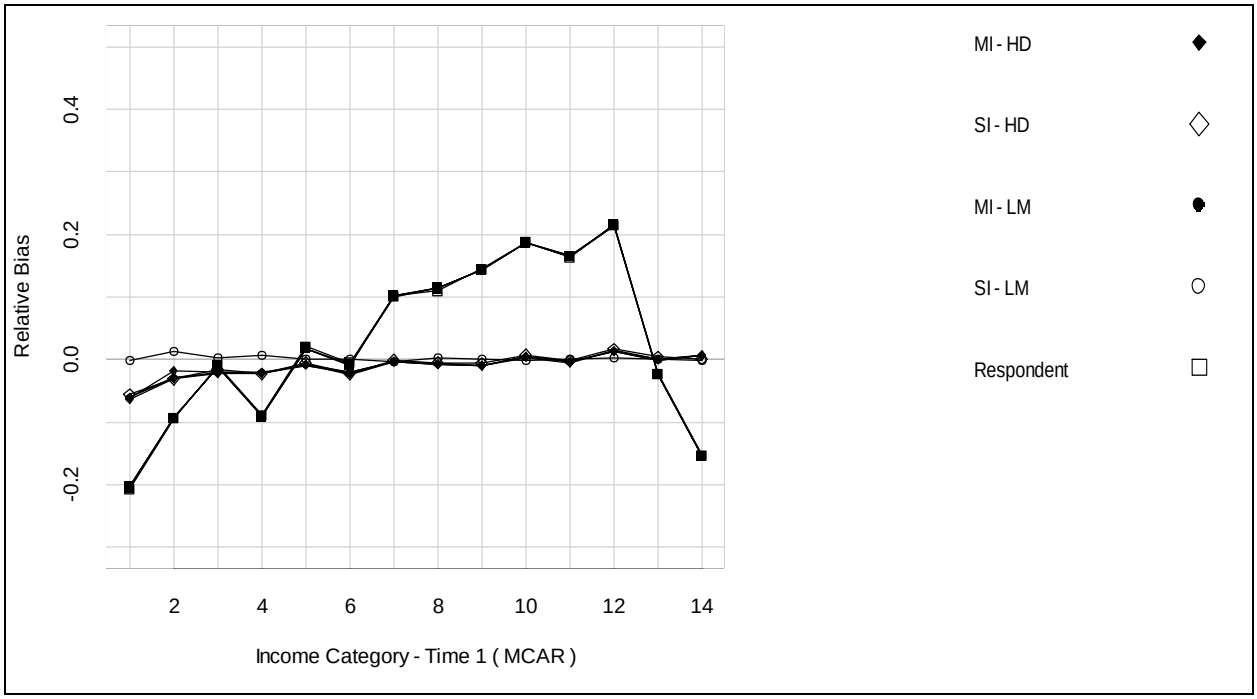


A.4.3 Relative Bias

The relative bias is the bias divided by the true values. The relative bias for MCAR is shown in ***Figure A.4-3***. Overall, there was very little relative bias for the respondents.

- For the SI-HD and MI-HD, there is a slight negative bias for the lower ordinal income categories.
- For the SI-LM and the MI-LM, there is a slight negative relative bias for the lower ordinal income categories that steadily increases up to the 12th ordinal income category.
- From the 12th to the 14th ordinal income category, for the single and MI-LMs, there is a steep decline resulting in a negative relative bias for ordinal income categories 13 and 14; this pattern was similar for the other time periods.

Figure A.4-3. Ordinal income variable category relative bias under missing completely at random (MCAR) missing data mechanism at the first time period



The relative bias for the MAR case is shown in ***Figure A.4-4***. In this figure:

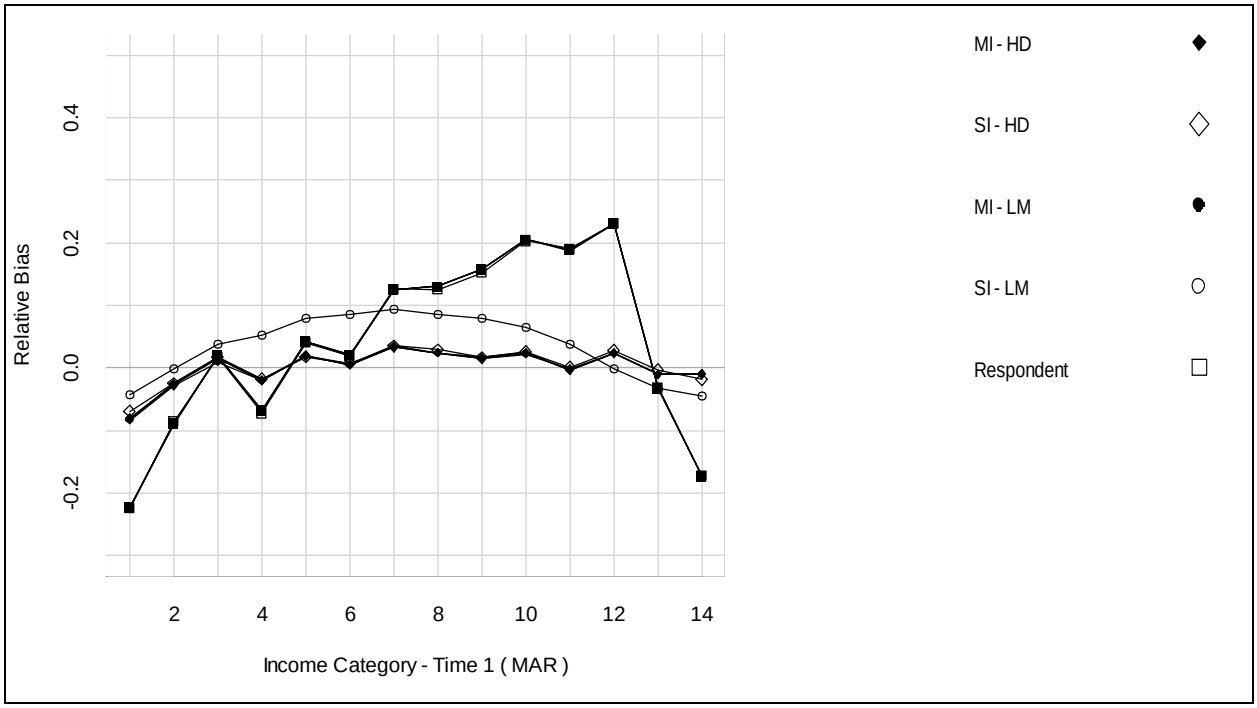
- For the respondent data, the relative bias is parabola-shaped, showing an increase from the 1st to the 7th ordinal income categories and a decrease from the 8th to the 14th ordinal income categories.

- There is a slight negative relative bias for the lower and upper ordinal income categories and a slight positive relative bias for the ordinal income categories in the middle for the SI-HD and MI-HD approaches.

- There is a slight negative relative bias for the lower ordinal income categories that steadily increases up to the 12th ordinal income category, which has a large positive relative bias for the SI-LM and the MI-LM.

- For the SI-LM and MI-LM, from the 12th to the 14th ordinal income category, there is a steep decline resulting in a negative relative bias for ordinal income categories 13 and 14, which was similar for the other time periods.

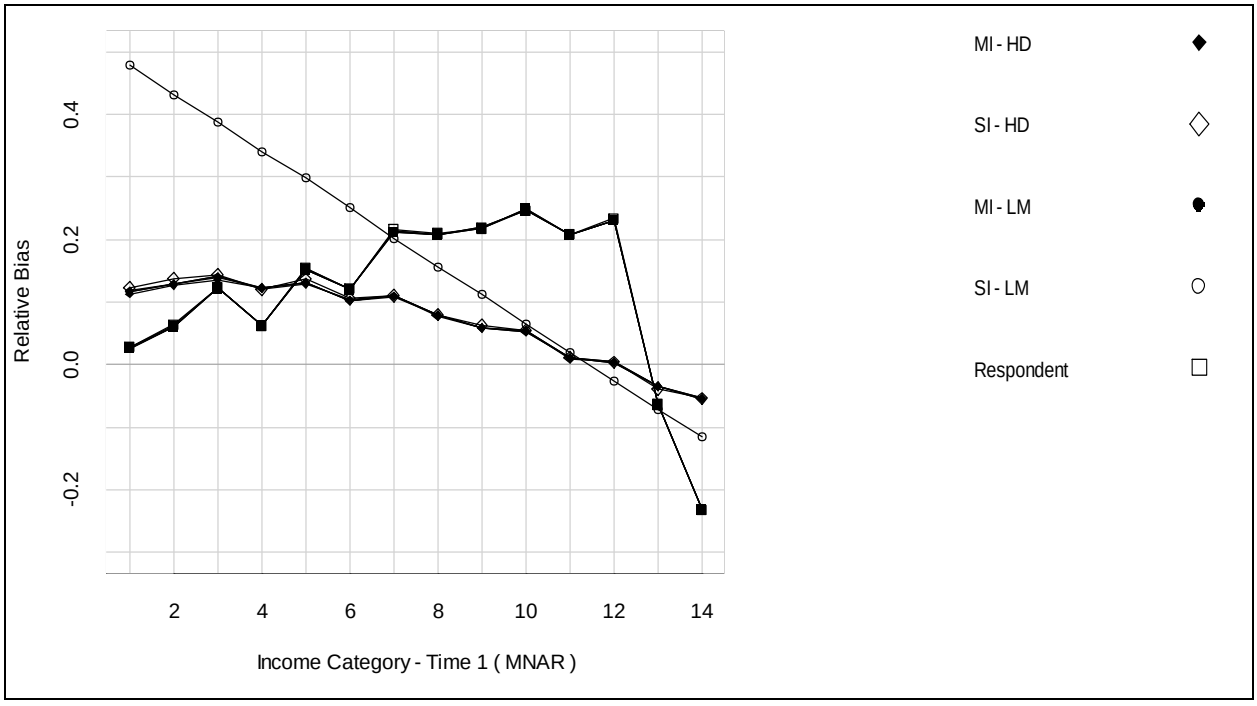
Figure A.4-4. Ordinal income variable category relative bias under missing at random (MAR) missing data mechanism for the first time period



For the MNAR case, the relative bias is shown in ***Figure A.4-5*** and indicates the following:

- For the respondent data, there is a large positive relative bias in the lower ordinal income categories with a steep decline to the upper ordinal income categories, resulting in negative relative bias for ordinal income categories 12, 13, and 14.
- For the SI-HD and the MI-HD methods, the pattern is the same as the respondents but is not as extreme particularly for the lower ordinal income categories.
- For the SI-LM and the MI-LM, there is positive relative bias for the lower ordinal income categories that steadily increases up to the 12th ordinal income category.
- For the SI-LM and the MI-LM, from the 12th to the 14th ordinal income category, there is a steep decline, resulting in a negative relative bias for ordinal income categories 13 and 14. The pattern was similar for the other time periods.

Figure A.4-5. Ordinal income variable category relative bias under missing not at random (MNAR) missing data mechanism

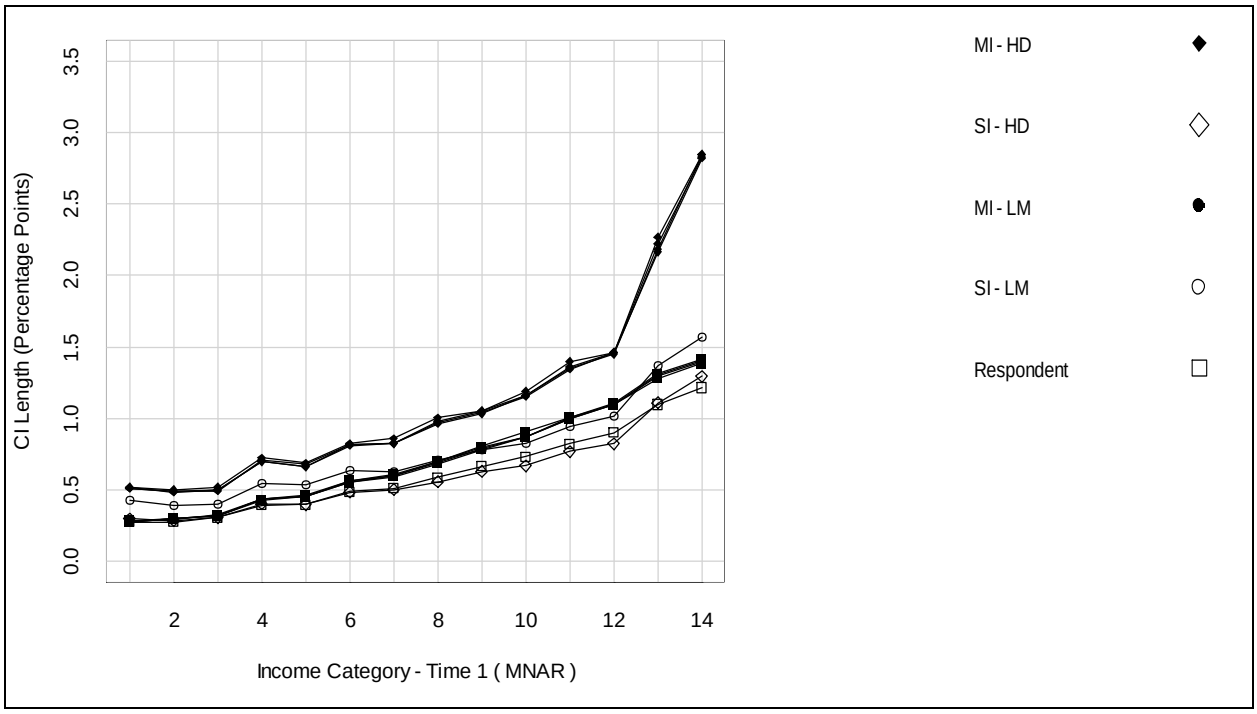


A.4.4 Confidence Interval Length

The figure for the 95% confidence interval length shows the length of the 95% confidence interval for the ordinal income categories. For all three missing data mechanisms and all four time periods, the pattern of confidence interval lengths was the same. ***Figure A.4-6*** presents the MNAR case for the first time period, which shows the following:

- The longest confidence interval length is for the MI-HD confidence intervals.
- For the MI-HD, the confidence interval lengths increase considerably for the ordinal income categories 13 and 14.
- The respondent and MI-LM confidence intervals are shorter.
- The shortest 95% confidence intervals are for the SI-HD and SI-LM approaches.

Figure A.4-6. Ordinal income variable category 95% confidence interval length under missing not at random (MNAR) missing data mechanism for the first time period

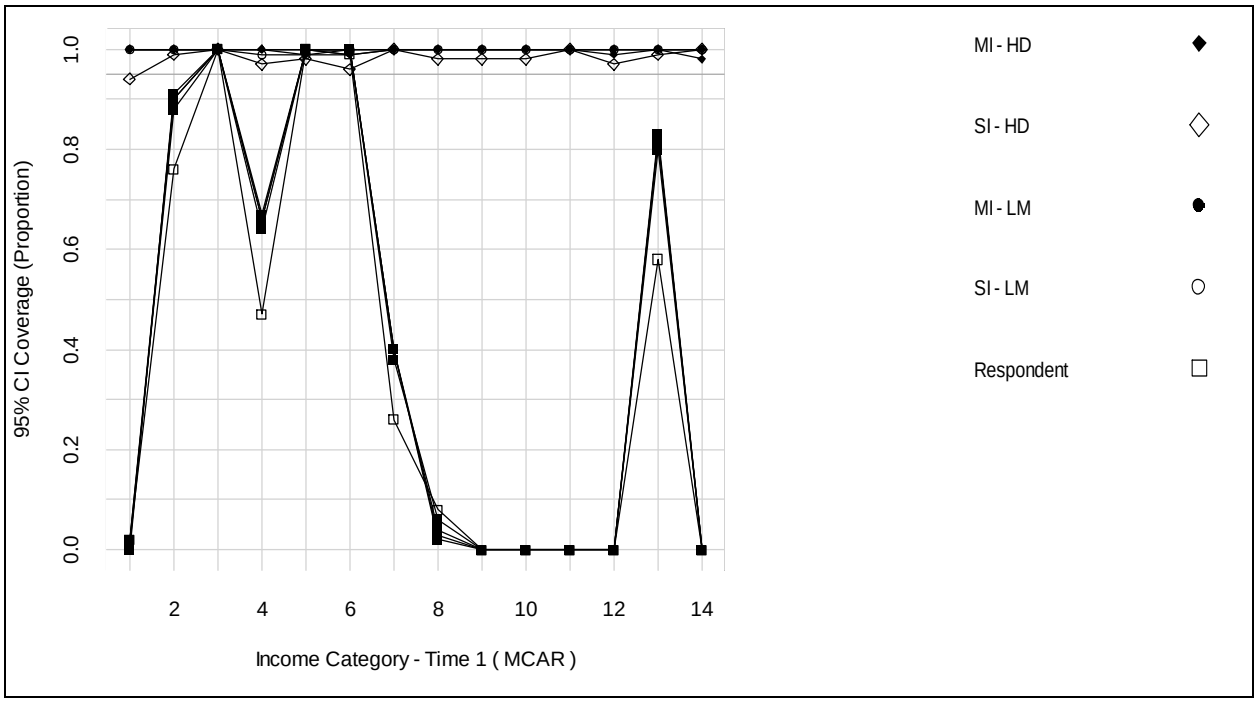


A.4.5 Coverage

As noted in ***Section A.2***, coverage was the primary evaluation criterion for evaluating the best imputation method. The figures in this section show the proportion of 95% confidence intervals that contain the true values for the ordinal income categories. The coverage proportions should be close to 0.95 for 95% confidence intervals. ***Figure A.4-7*** presents the MCAR results, which indicate the following:

- The respondent, SI-HD, and MI-HD 95% confidence intervals have good coverage proportions across all the ordinal income categories.
- The SI-LM and MI-LMs have poor coverage proportions for most ordinal income categories, particularly in the upper ordinal income categories.
- The other time periods have a similar coverage patterns.

Figure A.4-7. Ordinal income variable category 95% coverage proportion under missing completely at random (MCAR) missing data mechanism for the first time period



As shown in ***Figure A.4-8***, the MAR case shows the following:

- The MI-HD 95% confidence intervals have good coverage proportions across all the ordinal income categories.
- Generally, the respondent and SI-HD confidence intervals do not have good coverage proportions.
- The SI-LM and MI-LMs have poor coverage proportions for most ordinal income categories, particularly in the upper ordinal income categories.
- The other time periods have similar coverage patterns.

Figure A.4-8. Ordinal income variable category 95% coverage proportion under missing at random (MAR) missing data mechanism for the first time period

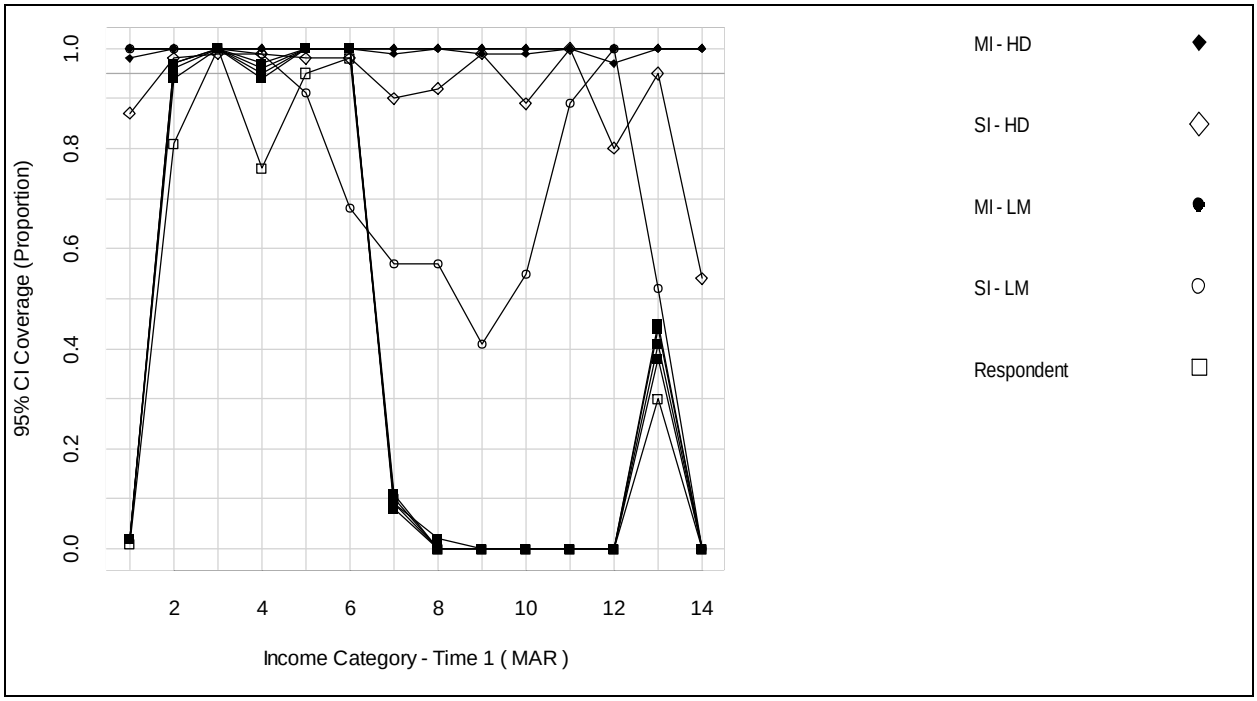
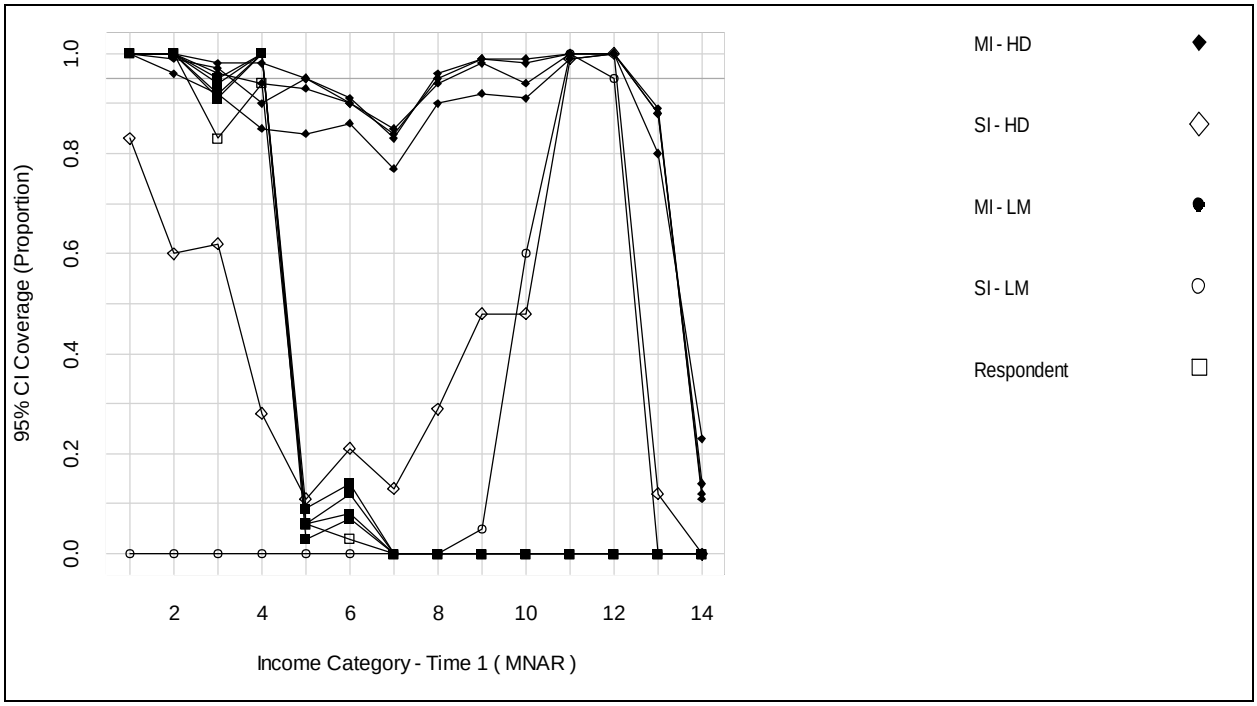


Figure A.4-9 shows the figure for the MNAR. As shown in the figure:

- The MI-HD 95% confidence intervals had good overall coverage proportions across the ordinal income categories, expect for the middle ordinal income variable categories and ordinal income category 14.
- The other coverage proportions—that is, respondent, SI-HD, SI-LM, and MI-LMs—had poor coverage proportions for virtually all of the ordinal income categories.
- The other time periods have similar coverage patterns.

Figure A.4-9. Ordinal income variable category 95% coverage proportion under missing not at random (MNAR) missing data mechanism for the first time period



A.5 Summary and Conclusions

The imputed estimates from the Monte Carlo simulation were very similar to the imputed estimates produced during the initial phase of research that used actual NCVS data. However, initial comparisons of two different imputation techniques—hot deck and linear models—were inconclusive because the true income value for the 33% of households that do not respond to the income question is unknown. This made it impossible to determine which imputed distribution of income from the two different imputation approaches was closer to the true income distribution. The Monte Carlo simulation created a population with properties identical to those of the NCVS population except that all household income values are known. Missingness can then be induced and the income values imputed, allowing the bias, relative bias, confidence interval length, and coverage of the imputation strategy to be known. For each imputation strategy, single and multiple imputation methods were tested. On the basis of the evaluation criteria, the following conclusions can be drawn:

- On the basis of the bias evaluation criterion and the estimates, the SI-LM and MI-LM approaches smoothed the ordinal categorical estimates so that the ordinal income categories 7–12 have higher estimates than they should and ordinal income category 14 has a lower estimate than it should.

- The coverage for both single imputation approaches and the MI-LM approach was poor because the confidence intervals were too short.

- The single imputation and MI-HD methods presented small bias and relative bias levels.

- The SI-HD method had good coverage when income was MAR, but it did not have good coverage when income was MNAR.

- The MI-HD method had good coverage regardless of the missing income pattern.

Based on these conclusions, it is clear that the hot deck strategy is better than the linear models strategy. Between the SI-HD and MI-HD approaches, the MI-HD approach performs the best because it has the best coverage proportions for this type of skewed ordinal variable. However, implementing a multiple imputation approach with NCVS data is likely to be too complex or difficult for the average user of public use NCVS data files. For this reason, the most appropriate method for imputing income data in the NCVS is the SI-HD approach.