NSAF

Methodology Reports

# 1997 NSAF Data Editing and Imputation

Report No.10

Prepared by:

Sarah Dipko
Michael Skinner
Nancy Vaden-Kiernan
John Coder
Esther Engstrom
Shruti Rajan
Fritz Scheuren

Assessing the New Federalism
*An Urban Institute Program to Assess Changing Social Policies*

# Preface

## Introduction

*1997 NSAF Data Editing and Imputation* is the tenth report in a series describing the methodology of the 1997 National Survey of America's Families (NSAF). The NSAF is part of the *Assessing the New Federalism Project* at the Urban Institute, in partnership with Child Trends. Data collection for the NSAF was conducted by Westat.

About the National Survey of America's Families (NSAF)

As discussed elsewhere (e.g., see especially Report No. 1 in the 1997 NSAF methodology series), NSAF is part of the Assessing the New Federalism Project at the Urban Institute, being done in partnership with Child Trends. Data collection for the NSAF was conducted by Westat.

In each round of NSAF, carried out so far, over 40,000 households were interviewed, yielding information on over 100,000 people. NSAF has focused on the economic, health, and social characteristics of children, adults under the age of 65, and their families. The sample is representative of the nation as a whole and of 13 states. Because of its large state sample sizes, NSAF has an unprecedented ability to measure differences between the 13 states it targeted.

About the 1997 and 1999 NSAF Methodology Series

The 1997 and 1999 methodology series of reports have been developed to provide readers with a detailed description of the methods employed to conduct the 1997 NSAF. The two series are nearly parallel, except for the documentation of the public use files, where an on-line system is being used for the 1999 survey and we are planning to reissue the 1997 files on a similar basis.

Report No 1 in the 1997 series introduces NSAF. Report Nos. 2 through 4 in both series—plus Report No. 14 in the 1997 series—describe the sample design, how survey results were estimated and how variances were calculated. Report Nos. 5 and 9 in each series describe the interviewing done in for the telephone (RDD) and in-person samples. Report Nos. 6 and 15 in the 1997 series and Report No. 6 in the 1999 series displays and discusses the comparisons we made to surveys that partially overlapped NSAF in content—including the Current Population Survey and the National Health Interview Survey, among others. Report Nos. 7 and 8 in both series cover what we know about nonresponse rates and nonresponse biases. Report No. 10 in both series covers the details of the survey processing, after the fieldwork was completed, including the imputation done for items that were missing. Report No. 11 in both series introduces the public use files made available.

In the 1997 series, there were additional reports on the public use files available in a PDF format as Report No. 13, 17-22. These will all eventually be superceded by the on-line data file codebook system that we are going to employ for the 1999 survey. The 1997 and 1999 NSAF questionnaires are available respectively as Report No. 12 in the 1997 series and Report No. 1 in the 1999 series. Report No. 16 for the 1997 series, the only report not so far mentioned contains

occasional papers of methodological interest given at professional meetings through 1999, regarding the NSAF work as it has progressed over the years since 1996 when the project began.

**About this 1997 Report**

Report No. 10 focuses on data editing techniques, including data processing, how data errors were dealt with, how edits were made, and coding guidelines, as well as how data were "filled in" when values were missing. Some discussion is also provided of the relative size of sampling variance to mean square error.

**For More Information**

For more information about the National Survey of America's Families, contact Assessing the New Federalism, Urban Institute, 2100 M Street, NW, Washington, DC 20037, telephone: (202) 261-5886, fax: (202) 293-1918, Website: **http://newfederalism.urban.org.**

Jenny Kenney
and
Fritz Scheuren

# Table of Contents

# Table of Contents (cont)



# List of Tables

## List of Tables (cont)



## List of Figures

# Chapter 1
# Overview of Data Editing and Item Imputation

## 1.1 Introduction

As discussed in Methodology Report No. 1 in this series, the 1997 National Survey of America's Families (NSAF) was a first-ever effort, put together to respond to major changes in welfare policy at the federal and state levels. The Urban Institute and Child Trends, who jointly led the effort, combined their expertise with the efforts of Westat, an experienced survey firm. Together, the ingredients were present for a successful endeavor. Still, there were many (mostly expected) difficulties, along with some unexpected developments.

This report is part of a methodology series that documents how the 1997 survey was fielded, treatment of survey methodology issues, and consequently, how the resulting data might best be analyzed.

The present report describes what was done to edit the NSAF data (see Chapter 2) and to impute for missing item information (Chapter 3). Some of the general analytic implications on the use of the resulting edited and imputed data are covered in the concluding chapter of this report (Chapter 4). There will also be further discussion of each survey variable in the public use files being prepared from the NSAF. Report No. 11 documents the first of these files—a file containing data items asked about children under 18 years of age. To give context to the NSAF effort, there is a bibliography at the end of this report. Included are some useful theoretical references plus citations to the editing and imputation practices employed in similar surveys (e.g., Welniak and Coder 1980, Schieber 1978, Oh and Scheuren 1980a and 1980b). The following sections give a more detailed description of what is discussed in Chapters 2 through 4 of this report.

## 1.2 Data Editing

As detailed in Chapter 2, the data editing process for the 1997 NSAF consisted of three main tasks: handling problem cases, reading and using interviewer comments to make data updates, and coding questions with open-ended (text strings) responses. Extensive quality control procedures were implemented to ensure accurate data editing.

Starting in January 1997, the NSAF editing staff began to categorize problem cases and review interviewer comments within the questionnaire. The review and categorization were the first steps in developing a data preparation protocol. The goal of the protocol was to allow for a systematic updating of information learned during the field period into the Computer Assisted Telephone Interviewing (CATI) data set. Once categories of problems/comments were defined, systematic "solutions" were attempted and their implementation set in motion.

Open-ended questions and "other-specify" responses were also coded. "Other-specify" responses were those in which a question had specific answer categories but also allowed for text to be typed into an "other" category. Open-ended questions had no precoded answer categories. Westat and the Urban Institute developed an interactive process for defining these codes. The structure for coding was reviewed and modified over several months. Coding was also done for questions that asked about occupation and industry. This coding was sent to the U.S. Census Bureau in order for the Bureau to apply its standard coding scheme.

Because data editing resulted in updates to the data, careful quality control procedures were implemented, both at Wetsuit and the Urban Institute. These measures involved limiting the number of staff who made updates, using flowcharts to diagram complex questionnaire sections, carefully checking updates, and conducting computer checks for inconsistencies or illogical patterns in the data.

## 1.3 Item Imputation

For most NSAF questions, item nonresponse rates were very low, often less than 1 percent (see Report No. 11 and Chapter 4). Basically, unless it was necessary to weight the sample (Report No. 3), items with very few missing entries were left unchanged. In no case were the answers to opinion questions imputed.

Still, there were important NSAF survey questions that were missing responses and were imputed in order to provide a complete set of data for certain analyses. For example, the determination of poverty status is crucial, but the income items that had to be obtained to make this determination were often incomplete. As is the case in many household surveys, the NSAF encountered significant levels of item nonresponse for questions regarding sensitive information such as income, mortgage amounts, health care decisions, etc. In fact, the income item nonresponse could range from 20 to 30 percent for these types of sensitive questions in the NSAF. Hence the problem needed to be addressed. For an introduction to the literature on this subject, see especially Kalton 1983, Little and Rubin 1987, and Lyberg and Kaspryzk 1997. The three volumes of *Incomplete Data in Sample Surveys* (Madow et al. 1983) are, in many ways, definitive.

As will be explained further in Chapter 3, the imputation of missing responses is intended to meet two goals. First, imputation makes the data easier to use. Second, imputation helps adjust for biases that may result from ignoring missing information, because characteristics of persons with nonresponses tend to differ from those of respondents.

The approach used to make the imputations for missing responses in the NSAF was "hot deck" imputation (e.g., Ford 1983). In hot deck imputation, the value reported by a respondent for a particular question is given or donated to a "similar" person who failed to respond to that

question. The hot deck approach to imputing missing values is the most common method used to assign values for missing responses in large-scale household surveys. For the NSAF, a hierarchical statistical matching hot deck design was used (Coder 1999).

The first step in this imputation process is the separation of the sample into two groups, those who provided a valid response to an item and those who did not. Next, a number of matching "keys" are derived based on information available for both respondents and nonrespondents. These matching keys vary according to the amount and detail of information used. One matching key represents the "highest" or most desirable match and is typically made up of the most detailed information. Another matching key is defined to be the "lowest" or least desirable. Additional matching keys are defined that fall somewhere between these two; when combined, these keys make up the matching hierarchy.

The matching of respondents and nonrespondents is undertaken based on each matching key. This process begins at the highest (most detailed) level and proceeds downward until each nonrespondent has been matched to a respondent. The response provided by the donor matching at the best (highest) level is assigned or donated to the nonrespondent. For the most part, respondents are chosen from the "pool of donors" without replacement. However, under some circumstances, the same respondent may be used to donate responses to more than one nonrespondent. By design, multiple use of donors is kept to a minimum. An imputation "flag" is provided for each variable handled by the imputation system. As will be seen in Chapter 4, imputation affects both bias and variance. For some analyses, researchers may wish to redo the imputations. This will always be possible with the public use files.

## 1.4 Analytic Concerns and Implications

The NSAF, by its very nature, was long and probing—often asking questions never brought together before in the same instrument. Clearly, the length and complexity of the questionnaire contributed to the challenges faced—notably by making the data editing more difficult and the amount of item nonresponse greater than it might otherwise have been.

Noninterviews have been extensively covered in Reports No. 7 and No. 8 in this Methodology Series. Report No. 8, includes information on how frequently nonresponse occurred at the screener or extended interview levels. Concerns about unit nonresponse bias were the topic of Report No. 7. Basically, while the unit nonresponse was sizable and raised the cost of the survey, there was little evidence of any serious overall bias after adjustment.

While the path through the evidence is different, we believe that the problems of data editing and item nonresponse are similar in their effect. Their existence in the NSAF created extra work, but after the adjustments described here, there is little evidence to suggest significant residual biases beyond those normal to household surveys. This point has already been made in Methodology

Report No. 1 and will be elaborated on further in Report No. 15, where the NSAF is compared directly with the Current Population Survey (CPS) and other major national samples.

Despite the high quality of data editing and imputation in the NSAF, researchers still need to be concerned with how they handle the resulting data. Hot deck imputation, for example, increases the sampling error (Scheuren 1976) in ways that are hard to measure without special procedures, like multiple imputation (Rubin 1987). We recommend that analysts of NSAF data that have been partially imputed employ an adjustment to their standard errors that corrects for the fact that they do not have as many cases as most variance estimation software (e.g., Wesvar) would assume. One conservative approach to this problem is covered in Chapter 4. References to other options are also provided.

Misclassification concerns arise in any editing or imputation procedure, unless the method of assignment perfectly places each missing or misreported case into the right group. A reweighting (or tailored reimputation) option might be employed by the final analyst rather than just using the general imputations provided. In the NSAF, therefore, we have provided a great deal of diagnostic information and audit trail details for the final analyst to do sensitivity analyses.

As with any data set, researchers will need to be on the lookout for anomalies. We believe, however, that these will be rare. Still, it is unlikely that we have been able to anticipate all the ways in which the data will be used. Almost certainly, there are errors that will be found when researchers carry out their detailed investigations. We would greatly appreciate being informed of any such discrepancies, so they can be brought to the attention of others. Further, depending on the nature of this information, new data sets may be made available.

# Chapter 2
# Data Editing

## 2.1 Introduction

This chapter discusses data editing for the 1997 NSAF. The data editing process consisted of three main tasks: handling problem cases (examined in Section 2), reading and using comments to make data updates (Section 3), and coding questions with text strings (Section 4). Additionally, quality control procedures were implemented to ensure that the data editing process was being performed accurately (Section 5).

Starting in January 1997, the data editing staff began to categorize cases that interviewers indicated were problems. All interviewers, as part of their interviewer training, had conducted practice interviews (see Report No. 9). A large pool of these interviews was thus created from the practice accounts; this pool became a starting point for the editing staff in their analysis of problem interviews.

Westat conducted coding for most "other-specify" and open-ended questions. "Other-specify" questions were those in which a question had specific answer categories but also allowed for text to be typed into an "other" category. Open-ended questions had no precoded answer categories. Westat and the Urban Institute developed an interactive process for assigning codes to the "other specify" responses. The structure for coding was reviewed and modified over several months. Coding was also done for questions that asked about occupation and industry. Occupation and industry items were sent to the U.S. Census Bureau so that its standard coding scheme could be applied.

Because data editing resulted in updates to the data, careful quality control procedures were implemented at both Westat and the Urban Institute. These involved limiting the number of staff who made updates, using flowcharts to diagram complex questionnaire sections, carefully checking updates, and conducting computer checks for inconsistencies or illogical patterns in the data.

Details of the data editing process are discussed in Sections 2.2 through 2.5, followed by a brief summary (in Section 2.6). Concerns and checks that researchers might make in analyzing the data are covered in Chapter 4.

## 2.2 Handling Problem Cases

One important task of the data editing staff was to handle cases identified as problems by the interviewers. In this section, the method that interviewers used to communicate problems is described, along with the system used by data editing staff to categorize and resolve them. One specific challenge related to the family definition is discussed in more detail. Finally, the total number of problem cases is provided, along with a summary of the process for problem resolution.

### 2.2.1 Communication of Problems

During data collection, an interviewer who experienced a problem while working a case could alert the project team in one of two ways. One method was to fill out a problem sheet on the case. Problem sheets from all the telephone facilities were sent to a single staff member at the Westat Telephone Research Center (TRC), who distributed them to the appropriate staff person for resolution. A log was kept of all problem sheets, and cases with similar problems were grouped to give an indication of volume for each discrete problem type.

The other method of communicating problems was to assign an electronic problem result code to the case, without manually completing a problem sheet. These cases were then reviewed electronically by a TRC supervisor and either refielded to the interviewers or distributed to the telephone center, programming, or project staff as appropriate for resolution.

### 2.2.2 Classification of Problem Sheets

The classification system for problem sheets was developed first in the practice account and then modified for the production account. As discussed above and in Report No. 9 of the Methodology Series, all NSAF interviewers began live interviewing work with a practice account. When their performance on practice cases was deemed to be of sufficient quality, they were moved to the production account.

#### 2.2.2.1 Use of Practice Account Problem Sheets to Categorize Early Problems

The practice account was used in tandem with the production account from January through April 1997. Separate problem sheet logs were kept for each account.

Many problems were resolved due to problems uncovered in the practice account—notably involving interfaces between the NSAF custom questionnaire program, standard Westat Computer-Assisted Telephone Interviewing (CATI) program modules and other external systems. The CATI programs for the practice and production accounts were identical; thus any program updates made to the practice account were also made to the production account. The categories used by the TRC to classify problem sheets are shown in Table 2.1 at the end of this chapter. Table 2.2, also found at the end of this chapter, shows the volume and relative size for each general category of practice account problem sheets.

#### 2.2.2.2 Production Account Problem Sheets

In the production account, interviewers worked on a greater number of extended interviews than they had in the practice account. This meant that the number of project decision problem sheets with extended interview issues was greater in production (Table 2.2) than in practice (Table 2.2). An additional category, manual queuing cases, was also added.

Manual queue cases (which were conducted only for the production account) created a new category of problem sheets that accounted for nearly one-quarter of all production problems. Cases handled in the manual queue were those that needed individualized attention. Examples included situations where respondents moved and had to be asked about their original household; also included were respondents not mentally capable of answering the questionnaire (who had to be interviewed by proxy). See the Telephone Survey Methods Report (No. 9 in this Methodology Series) for a more detailed discussion of the manual queue activities.

### 2.2.3 Problems Recorded in the Electronic Queue

Cases that were coded problems via a problem result code (a code of 8) began to accumulate in January for the screener questionnaire and in February for the extended questionnaire. In Table 2.4, the distribution of electronically coded problem cases is categorized by screener and extended interviews. Problems coded with the initial code of 8 are shown along with three subcodes for special queues to which these 8's could be assigned for review. These subcodes, 81, 82, and 83, are all forms of problem code 8 and were used to indicate who was responsible for investigating the case further—TRC supervisory staff, statistical staff, or programmers, respectively.

### 2.2.4 Problems Related to Family Definition

In fewer than 150 households, problems occurred with the extended interview for a variety of reasons related to the family definition. Although these problems affected a small percentage of cases, they required a great deal of time to resolve and resulted in important changes to the questionnaire. Thus, these changes are discussed in some detail.

The definition of a family in the NSAF was designed to include all persons who could be considered family by respondents and who would likely share income and food in the home. Households had the potential to include multiple families if unrelated persons lived in the same household and did not fit any of the categories for inclusion in the family that are outlined below.

The family definition for the NSAF went through two sets of changes during the course of data collection. The changes were made to accommodate less common family configurations that resulted in

problems with how extended interview information was collected. According to the original definition of family for the NSAF, the family included all persons living together in a household who were related by blood, adoption, or marriage to the target person (either the selected child or adult). Also included as family were any unmarried partner that the respondent may have had, plus the children of the unmarried partner. The relatives of an unmarried partner were originally not included as family; however, they were included at a later point in data collection.

When data collection began, it became obvious that several types of exclusions from the definition of family had caused persons to be counted on different family lists for different interviews, depending on which household member was the selected person. In households with two or more interviews, this resulted in "overlapping" families. It was also noticed that in some households with only one interview, some family members were missed entirely. Some of the problems were due to coding errors (e.g., coding a stepfather as a stepson by mistake), but most were due to exclusions from the family definition. Coding errors represented about 10 percent of the total overlap cases.

In early March 1997, an initial adjustment in the family definition was made to accommodate several problems in the way that families were classified in the interview. One problem had to do with respondents telling interviewers that they did not know (or refused to indicate) the relationship between one person in the household and another. Unknown relationships could cause the person or persons in question to be excluded from the family list. It was decided that if none of the relationships for a person was known, that person would be included in the first family that was defined. Although this may have been inaccurate in some cases, it was necessary to include the person in some way in the interview. It was more likely than not that the person belonged to a family in the household, but for some reason the respondent was reluctant to give out information about this person.

Another problem with the family definition was sometimes caused when a person was said to be an "other relative" of another household member, but the respondent did not tell us what type of relative the person was. The list of categories in the question about type of relative included all the possible "other relative" codes, such as mother/father-in-law, aunt/uncle, niece/nephew, etc. The relative thus was not included as a family member. A change was made to include an "other" category in this question so that these relatives would be counted as family members.

In addition, it was decided to employ a lead-in question that asked about the relationship of each person to each other person in the household as the basis for forming families (see the full 1997 questionnaire in Report No. 12, Section D). This allowed persons to be included in the family for whom "other relative" relationships were not reported in the question that asked about type of relative.

Finally, there were problems because relatives of unmarried partners of the most knowledgeable adult (MKA) were not included. This was causing persons to be counted in two different families because unmarried partners were included in the MKA's family and again in a separate family with their relatives. To avoid collecting information about the same person in two different interviews, it was decided to include all relatives of the unmarried partner of the MKA in the MKA's family.

Although these changes resolved the majority of the issues surrounding the family definition, by April it was discovered that a few more coding accommodations were needed for families to be defined correctly. The following steps were taken to address the remaining inconsistencies.

First, an edit was put into the program to indicate early in the interview when a family definition problem had arisen. In these cases, interviewers reported that they had a problem with their computer and made an appointment to call the respondent back. Once the family definition problem was fixed, the interviewer called back to complete the interview. In households with only one interview, if the family definition "missed" an individual, cases could only be stopped if the respondent expressed concern over the family list (which was read to the respondent) and the interviewer coded the case a problem. Problems in "one-interview households" generally involved children who were missed.

As a second step in addressing family definition issues, problem cases were categorized to determine what types of exclusions from the family definition were causing the most problems. Most of the overlapping family problems were due to the following exclusions:

- Exclusion of unmarried partners (of focal children and persons not selected for an interview) and their relatives

- Exclusion of the MKA's spouse when the child was coded as a nonrelative of the spouse (i.e., rather than as a stepchild)

- Exclusion of the focal child's relatives or nonselected person's spouse and the spouse's relatives was also an issue because the program was not originally counting these persons as family

- Exclusion of relatives of MKA's (e.g., the MKA's sister) when a focal child was coded as "foster child" of the MKA, if the relationship between the relative and foster child was reported as "other nonrelative"

- Exclusion of children (focal and nonfocal) who were "other non-relatives" of selected respondents (e.g., foster siblings, children who share an MKA with a selected child)

- Exclusion of children who were "other nonrelatives" of all household members (e.g., the child was unrelated to anyone in the household and was NOT a sampled child. For adults, this would have been considered a separate family, and it would be reasonable to leave the adult off the first family list. For children, this presented more of a problem because presumably someone in the family was responsible for the child, and the child should have been on a family list rather than form his or her own family).

When the persons involved in the exclusions were included as family members in the program, there were only rare cases of problems with overlapping families. Remaining family definition problems were due to unusual relationship combinations and were fixed manually. Because the program handled the

vast majority of relationship combinations, no more programmatic changes were made to CATI. The program specifications for the revised family definition are contained in the CATI specifications for Section D of the NSAF questionnaire.

Cleanup of the cases that had overlapping family problems or missing persons were handled in two different ways. For cases that were not complete and were not far along in the interview, the relationship codes for the household were examined, corrected as appropriate, and the interview restarted with the correct family list. For most of the cases that were completes (53 households), the data were fixed or reentered by using answers from first and subsequent interviews to properly resolve the family. Some completed cases did not have adequate information to rebuild the case. These cases were coded as nonresponse. In total, 43 of the 53 households that had overlapping family problems at the end of data collection were fixed.

Because the family definition changed over time, there were cases in which the family was defined according to previous versions of the family definition, rather than the final one. There were 91 cases in which the family definition would have differed had the households been called after the final family definition went into effect.

One category of definition problems within the 91 cases developed early in data collection and occurred when persons who were related were reported to be nonrelatives by respondents. For example, if there was a child interview in the household and the male spouse of the MKA was coded as not related to the focal child (even though the real relationship would be stepfather), the spouse was excluded from the family. For these interviews, the spouse questions were asked throughout the extended interview for the stepfather (e.g., employment data and education data), but the family questions (e.g., health insurance data and family income data) were not. There were also a variety of other categories within the 91 cases that were all related to a change in the family definition.

### 2.2.5 Overall Number of Problem Cases

The total number of problem cases is shown in Table 2.5. Both problem sheets and problems in the electronic queue are described. Including practice account cases, a total of 17,791 problems were handled.

### 2.2.6 Problem Resolutions

In March, the categories of problem cases that had accumulated in the problem queue were summarized and recommendations made on their resolution to the Urban Institute. The most difficult cases would be handled by working the cases in the manual queue (see 2.2.2.2 for a description of the manual queue). The manual queue work began in mid-May after approval for these recommendations was received from the Urban Institute.

One specific category of problems—enumeration errors—took longer than the others to resolve. These problems touched upon sampling issues and, therefore, required careful review and treatment in order to preserve the integrity of the study's sampling procedures. Westat summarized these problems and proposed recommendations to the Urban Institute starting in early May. By early June the cases with these so-called “enumeration errors” were being corrected.

For some problems, no database updates were necessary and the cases could simply be released for general interviewing with a message; for example, suppose a case was stopped in the middle of the health insurance section of the extended interview, and the problem message typed by the interviewer read: "respondent remembered that their health insurance is a private plan, and not from their employer—can't back up to change the answer." The solution to this was to release the case with a message stating, "Case will restart in Section E; re-ask the insurance questions."

Similarly, problems related to miscoding responses in the screener questionnaire generally did not require database updates. This was because every time the screener was reopened it restarted at the beginning and re-asked every question. As long as the screener was assigned a problem result code while still an interim—not completed—screener, the database would be reset when the questionnaire was reopened.

Only the cases moved by the TRC into the special review queues were investigated for possible updates. When cases were changed, they were often re-released for general interviewing. Some examples of the types of cases reviewed by project staff were those in which an error was made in enumerating a household member or when the person named as most knowledgeable about the child had to be changed.

Other types of problems required special interviewer handling, even after changes were made to the database. For example, as discussed in Section 2.4, in some cases respondents misreported relationships between household members, resulting in problems with the family construction for interim extended interviews. These problems were moved to the manual queue.

The manual queue interviewers served, in a sense, as an auxiliary branch of the data editing team. They were trained to recognize a wide variety of problem situations, many involving enumeration errors, and to proceed slowly to avoid repeating the error. They were also given the authority to offer small monetary incentives to gain cooperation from proxies or facilitators when the selected respondent was unable to complete the interview. This team of interviewers also conducted data retrieval for a set of cases in which respondents who had previously completed an interview had omitted their spouse or partner from the roster and were offered $25 to redo the entire interview correctly.

## 2.3 Handling Interviewer Comments

Another important data editing task was reading and using interviewer comments. Comments are text phrases that interviewers type in special entry windows in CATI when the respondent makes a

statement that the interviewer wants to record but is unable to enter as a standard response in the questionnaire. Sometimes these phrases are merely an elaboration on a previously recorded response or an expression of opinion. Other times, they indicate that an update needs to be made.

### 2.3.1 Training

Once data collection began and completed interviews were being generated, a separate data editing staff was trained to "clean" the completed interviews by reviewing interviewer comments and coding other-specify and open-ended questions (coding is described in Section 4).

Training consisted of two stages. The first stage was familiarization with the survey materials. Staff members read the interview manual and the specification manual to introduce them to the questionnaire. They also used the test account to run mock interviews to see how the skip patterns changed based on the characteristics of the respondents.

The second stage of training involved practice on the tasks that would be performed. The data editing supervisor provided examples of comment review and other-specify coding and checked 100 percent of their work until they were sufficiently comfortable with the material. After that, the supervisor continued to check between 60 percent and 80 percent of their work.

### 2.3.2 Comment Review

During data collection, a list of comments was generated daily for all the cases that had comments the previous day. These comments came from both complete and incomplete cases. Sections B, C, F, L, M, N, O, and P of the questionnaire were eligible for comment review.

Initially, only those from completes that made it through the end of the interview (Section P) were considered for updates. Later, comments were also reviewed for “partial completes.” At the end of the data collection, the Urban Institute established a definition for a “partial complete,” which was an interview that had progressed at least through the end of Section K. It might be worth noting that the definition of a “partial complete” may be changed in the 1999 NSAF, since the item imputation software proved so successful (see Chapter 3). Usually, it is better to adjust for item nonresponse than to adjust for unit nonresponse and this should prove true in the NSAF.

The comment review process had four steps: counting, recoding, checking, and entering. Counts were kept on the number of complete cases with comments, the number of comments in complete cases, the number of incomplete cases, and the number of comments in incomplete cases. Further counts were kept of the number of comments eligible for updating, the number of updates, and the number of problems marked for later review. In order to project staff needs and track production output, all comments for complete cases were categorized as shown in Table 2.6.

The classifications were loosely made based on the answer and the comment, without further investigation. After counting, each comment from an eligible section of a completed case was investigated. The interview was updated if the information in the comment was clearly more accurate than that in the interview.

Comments labeled "HH" (household), income, sex, and age were not reviewed as part of normal data editing comment review. However, those associated with problem cases and those found in text string searches of comments were used to aid in the resolution of problems found at the end of the data editing process via a set of systematic checks for unlikely relationship, age, and gender roster entries.

### 2.3.3 Use of Comments from Complete Cases to Help with Interim Case Problems

During production, the data editing staff notified the project team about any sections or questions of the questionnaire that were generating a large number of comments. There were three areas that received numerous comments and resulted in changes to questions or procedures.

One section that received many comments was Section J. In Section J, respondents were asked about nonsalary forms of income, including such things as annuities, pensions, interest, dividends, and rental property income. If respondents indicated that someone in the family received a certain type of income, they were asked to identify the recipient and report the amount of income received from that source during the previous calendar year. There were many comments in this section of the database indicating that respondents were reporting information about joint accounts. In some instances, they would identify two recipients and report the same income for both people, and then make a comment that "this is a joint account." It was not clear, however, if the total amount received was split between the two recipients or if it was reported twice, thus falsely inflating the family income.

The project staff consulted with the Urban Institute about how to handle this issue and decided upon two approaches. First, the questions in Section J that asked, "Who received this source" were expanded by adding a sentence, "Please give me only one name if two or more people shared income from the same pension or annuity." Second, a training memo was distributed to all of the interviewers explaining the issue with joint accounts, why this was a problem, and what the revised questions looked like. Instructions were provided to help interviewers write comments that would clarify, rather than confuse, the values recorded in the database.

Two other issues were also evident from a number of comments. One involved reports from respondents about the age of siblings of the focal child who lived outside the household. The question asked about siblings under the age of 18. After some respondents answered that the focal child had siblings outside the household *under* 18 years old, they sometimes later told the interviewer that the sibling was *over* 18 years old. Interviewers were asked to clarify these statements with respondents so that the comments about these questions would be clear.

Interviewers were also asked to clarify information in comments dealing with annual income. If a respondent could not report total annual income, he or she often gave the amount earned on a weekly basis. In these cases, the interviewers were asked to find out how many weeks or months the respondent had received that income, so that the data editing staff could calculate an annual income based on the comment.

## 2.4 Coding Questions with Open-ended or Text String Responses

Coding for the 1997 NSAF included other-specify and open-ended questions. As stated previously, other-specify questions were those in which a question had some specific answer categories but also allowed for text to be typed into an "other" category. Open-ended questions had no precoded answer categories. Westat and the Urban Institute developed a highly interactive process for dealing with these types of questions.

### 2.4.1 Other-Specify and Open-ended Question Coding

For other-specify and open-ended questions, a frequency listing of the first 100 to 1,000 responses was made. A scheme was determined for upcoding to the original options and for the creation of new codes to handle the new responses for other-specify questions. The Urban Institute built on this scheme, modifying it somewhat, and assigned a list of codes and rules for using them. The data editing team used the list of codes and the related rules, to do the coding. For many of the questions, the Urban Institute reviewed the final coding, and the data editing staff made any necessary changes.

In July 1997, a partial data delivery was made to the Urban Institute. Upon review of the delivery files, the Urban Institute refined codes to many of the other-specify questions and provided more detailed instructions for the data editing staff. This required a review of all of the updates for most of the other-specify and open-ended questions. In February 1998, the Urban Institute provided a series of requests for changes and corrections. These requests led to changes for some questions in which updates required modifications to skip patterns.

Variable-by-variable information on “other-specify” and open-ended coding will be found in the public use file codebooks (e.g., Report Nos. 11 and 13).

### 2.4.2 Occupation and Industry Questions

There were three questions about occupation and industry in the questionnaire. The first asked about the kind of industry the person worked in for his or her main job. Another question asked whether the business or organization was mainly manufacturing or something else. A third question asked what kind of work the person did, that is, what his or her occupation was.

Because the codes for the occupation and industry questions needed to be consistent with coding conducted by the U.S. Census Bureau; these items were sent to Census for coding. Once Census returned the codes, they were posted to the database.

## 2.5 Achieving Data Consistency and Quality

Five of the primary methods used to ensure accurate data editing are discussed here. In various ways these methods were used at both the Urban Institute and at Westat. Specifically, an attempt was made to limit the staff who could make updates, flowcharts of the complex question dependencies were employed extensively, and there was regular consultation between Westat and the Urban Institute when issues arose in complex updates. Finally, updates by Westat were checked by computer and subsequently cross-checked by the Urban Institute. Naturally, some modifications to the data were required after review. New procedures were implemented to further prevent errors in the updating process.

### 2.5.1 Limiting the Number of Staff Who Made Updates

In order to avoid contaminating the data, the number of people making updates to the database was severely limited. When Urban Institute or Westat employees found updates that were needed, these updates were made only by data editing staff who were familiar with the questionnaire and had been trained specifically for the NSAF.

### 2.5.2 Using Flowcharts

The data editing staff created flowcharts of the CATI specifications for the sections that were eligible for comment review. These flowcharts assisted in making downstream edits to other sections or questions affected by the data changes. Many of the sections were quite complicated, with skip patterns affected by numerous factors. Because of this, it was difficult for the staff to make consistent, comprehensive edits without these diagrams. These flowcharts have not been reproduced for researchers; instead, the full 1997 NSAF questionnaire has been provided as part of Report No. 12 in this Methodology Series.

### 2.5.3 Consulting with the Urban Institute

The data editing staff initially restricted updating to comments from relatively self-contained sections of the questionnaire (those that did not affect other questions or skip patterns) and to other-specify responses. When the data editing staff worked on sections that were not as self-contained, it worked with the Urban Institute to identify the questions that could be updated and flagged the sections that had been changed so that analysts would be aware of them. This interactive process was essential to making sure that data editing was performed correctly across sections.

### 2.5.4 Checking Updates

Updates were checked at two stages. In the beginning of data collection, the procedure for making updates was to write the intended updates first on a hard-copy printout. This printout was checked for accuracy and logical effects on other questions or skip patterns in the questionnaire by a data editing supervisor. The updates were then entered onto the computer and checked again. The second check focused on looking for duplicate entries and missing values.

After initial delivery of the data files, the Urban Institute asked for corrections to some of the updates. Consequently, the procedure for making updates was changed somewhat to further prevent errors. The text responses, existing values and corrected values were all listed on a spreadsheet. The responses were checked (and in some cases altered) by the data editing supervisor on the spreadsheet. The spreadsheet was then converted into a set of case update records automatically, ensuring that the updates would run exactly as intended on the spreadsheet.

### 2.5.5 Performing Programmatic Checks

Westat also worked with the Urban Institute to clean the database by performing a series of programmatic checks in order to find inconsistencies or illogical patterns. The problems stemmed from errors during the interview, such as misreporting or miscoding of relationships or ages, or from changes made by the data editing staff, such as updating one, but not all, of the variables that were dependent on other variables in the interview.

This process of data cleaning was highly interactive, as suggestions for specific checks came from both Westat and the Urban Institute. It was also an iterative process that ran for a period of 6 months, starting at the close of data collection, with the results of some checks suggesting new or expanded checks. Following each delivery of database files to the Urban Institute, programmatic checks were run at the Institute and results were communicated to Westat. Parallel checks were then run at Westat, problem cases were identified, and the errors were resolved prior to the next database delivery.

For errors caused by incomplete or incorrect edits to variables that were dependent on other variables, Westat questionnaire programmers developed edit check programs to identify cases and repair the errors. These edit check programs were primarily "forward" checks, that is, they verified that a response was coded for a questionnaire item when expected. The programs could not, however, identify responses left in questionnaire items that should have been deleted—the "reverse" check. The Urban Institute's data checking programs did, however, identify these types of problems, which provided valuable assistance to the data checking team in its effort to produce a clean database. In all, Westat programmers updated approximately 5,000 database values in over 2,600 households during this phase of database cleaning.

Other types of errors in which variables such as relationships, age, or gender were involved were identified programmatically but reviewed and corrected by hand. A team of three Westat data coders reviewed and corrected the initial set of problem cases, which included errors such as male mothers, female fathers, children older than their parents, and children with two mothers or two fathers (excluding same-sex spouse/partner situations). During this phase of database cleaning, approximately 300 households with data errors were reviewed and edited.

After this initial set of cases was resolved, a single Westat data analyst reviewed other problems of this type. This analyst checked and edited these cases over a period of 5 months and developed a set of quality control procedures to ensure that the best solution was found for each individual case. The errors reviewed included illogical or impossible items such as multiple spouses, children older than parents, grandchildren older than grandparents, age errors, and gender errors. Also, all same-sex spouse/partners were checked and many turned out to be invalid because they were caused by a miscoded sex for one spouse or partner.

Prior to making updates, each case was checked to see if any previous updates had been made to the age, sex, or relationship variables involved in the error. In addition, a program was run to produce information sheets on each case showing the relationships between all household members. Family structure diagrams were drawn on these sheets, inconsistencies were identified, and necessary edits were noted. For cases that were difficult to resolve, other information was reviewed, such as interviewer comments, messages, and problem descriptions, in order to ensure proper handling of the error.

Cases with similar errors were reviewed together, and following the review the cases were all updated at one time. This ensured consistency in the handling of a discrete problem type. Following each wave of updates, a program was run to check for the full "set" of errors that had been identified at that point in time to ensure that data editing had not created any new errors. Approximately 1,000 households were reviewed and edited during this phase of database cleaning.

## 2.6 Summary

Westat conducted the majority of the data preparation steps for the 1997 NSAF. Initially the Urban Institute's role was just to provide specifications and then run checks on the completed work. As time went on and the Institute became more familiar with the data structures being employed by Westat, it began to assume a larger role, culminating with the Institute's preparation of the public use files from the survey data.

Even though the Institute's role gradually changed, the handling of problem interviews and interviewer comments was always the province of Westat. As we have seen, problem cases were communicated to the Westat data editing staff in two ways. The first way was through problem sheets. The total number of problem sheets in the production account was 3,923. The second method of communicating

problems was for interviewers to code a case electronically. The total number of problems coded in this way was 13,476—about 5,000 of which required special follow-ups.

To summarize, Westat first categorized problem cases by examining those it found in the practice account. Many problems with the scheduler and other systems were worked out in this account and corrected for production. However, the volume of problems related to the extended interview was much greater in production. This was because of the greater number of interviews conducted and because more investigative work was done to correct data. Some of these corrections required individualized handling by interviewers and resulted in the creation of a new category of cases that were only handled by specially trained interviewers. One type of problem that was particularly difficult to solve involved cases in which the listing of household members and/or relationships among them were incorrectly reported or entered. The development of procedures for problem resolutions began in early March 1997 and continued through June 1997. Some of the household problems had to be later handled again by the Urban Institute because they were so complex.

Because 1997 was the first round of NSAF, interviewer comments played a major role in making its updates to the data in some sections of the questionnaire. Lists of comments were generated daily and were put into 10 categories. Comments within some categories (e.g., those about household composition) were not used for updates but did help resolve problem cases. Other comments were used to give feedback to interviewers about the types of issues to clarify with respondents during future interviews.

The Urban Institute and Westat worked together to implement coding schemes for other-specify and open-ended questions. Westat made initial recommendations and the Urban Institute made decisions about codes and rules. The coding classifications were reviewed and modified in an iterative process over several months. The one major exception to this is that open-ended questions about industry and occupation were sent to the U.S. Census Bureau for coding. This was done so that data for these questions would be comparable to other databases that used the Census coding scheme.

The Urban Institute reviewed all updates after they were made. Despite consulting with the Urban Institute before complicated updates were made and then double-checking them, the survey data structure is so complicated that inconsistent or improbable data were subsequently found. As such, we caution NSAF data users that they may find minor remaining inconsistencies when using the public use files. We do expect, though, that these inconsistencies will be infrequent.

**Table 2.1 — Categories Used to Classify Problem Sheets**

| Category | Description |
|---|---|
| Correspondence | Respondent requested study results or letter |
| Database | Required a database change so the case could be worked (e.g., male respondent was coded as a female and incorrect questions were being asked of him.) |
| Programming | Other programming problem including extended questionnaire |
| Project decision | Required a decision by project staff |
| Scheduler | Problem during case delivery |
| Other systems: | |
| Mainline | Problem within screener or contact screens |
| Telephone | Problem with voice telephone line |
| Utility | Problems with utilities that looked at problem cases, moved result codes, found ID numbers through telephone numbers, etc. |
| VAX/SYS | Problems with hardware or systems software. Included data communication lines. |

**Table 2.2 — Practice Account Problem Sheet Summary**

| Category | Total Sheets | Percentage |
|---|---|---|
| Correspondence | 62 | 16 |
| Database | 27 | 7 |
| Programming | 115 | 29 |
| Project decision | 19 | 5 |
| Scheduler | 60 | 15 |
| Other systems | 109 | 28 |
| Total problem sheets | 392 | 100 |

**Table 2.3 — Production Account Problem Sheet Summary**

| Category | Total sheets | Percentage |
|---|---|---|
| Correspondence | 997 | 25 |
| Database | 474 | 12 |
| Programming | 862 | 22 |
| Project decision | 319 | 8 |
| Manual queue | 889 | 23 |
| Scheduler | 203 | 5 |
| Other systems | 179 | 5 |
| Total Problem Sheets | 3,923 | 100 |

**Table 2.4 — Problem Case Distribution**

| Problem Queue | | Questionnaire | | Total Problems |
|---|---|---|---|---|
| Code | Description | Screener | Extended | |
| 8 | Initial problem queue | 3,092 | 10,384 | 13,476 |
| 81 | TRC review queue | 564 | 1,999 | 2,563 |
| 82 | Project staff review queue | 84 | 1,921 | 2,005 |
| 83 | Programmer review queue | 136 | 348 | 484 |
| 81–83 | Sum of special queues | 784 | 4,268 | 5,052 |

**Table 2.5 — Total Number of Problem Cases**

| Source | Number of Problems |
|---|---|
| Problem 8's (electronic) | 13,476 |
| Production account problem sheets | 3,923 |
| Practice account problem sheets | 392 |
| Total | 17,791 |

**Table 2.6 — Categorized Comments for Complete Cases**

| Category | Definition | Example |
|---|---|---|
| HH (household) | Household information is incorrect | R is 45 and is in HH |
| Inc | Income information in comment | R paid by the hour not salaried |
| Language | Problems understanding interviewer | R speaks mostly Spanish |
| Address/Name/Sex | Updated information to HH member's address, name, or sex | R is male not female |
| -8 Poss.Up | Current code is don't know (-8); possible update, but more checking needed | Had Medicare, Medicaid |
| -8 Up | Current code is -8; update needed | Went to doctor 4 times not 5 |
| -8 No | Current code is -8; update not needed | Doesn't know that information |
| Non-8 Poss.Up | Current code not -8, possible update | Insurance through union |
| Non-8 Up | Current code not -8, update needed | All the time, not some of the time |
| Non-8 No | Current code not -8, update not needed | Not sure, but thinks so |

# Chapter 3
# Data Imputations

## 3.1 Introduction

As is the case with any household survey (e.g., the Current Population Survey or CPS), both unit and item nonresponse can and usually do arise. This chapter is about what was done in the presence of NSAF item nonresponse, but it might be worthwhile to open our discussion by making a few remarks about unit or noninterview nonresponse.

Noninterviews have been extensively covered in two other reports in this Methodology Series: Reports No. 7 and No. 8. In Report No. 7, information was included on the frequency of nonresponse at the screener or extended interview levels. Concerns about the unit nonresponse bias were the topic of Report No. 8. Basically, while the nonresponse was sizable and raised the cost of the survey, there was little evidence of serious overall bias after adjustment.

While the evidence is different, we believe that the problem of item nonresponse is similar in its effect. The existence of item nonresponse made for much extra work but after the adjustments described here, there appear to be no residual biases—beyond those normal to household surveys (e.g., the understatement of incomes). This observation has been made in Methodology Report No. 1 and will be elaborated on further in Report No. 15, where the NSAF is compared directly with the CPS and other major national samples.

For most questions the item nonresponse rates were very low, often less than one percent (see, e.g., Report No. 11). Basically, unless needed to weight the sample (see Report No. 3), items with very few missing entries were left unchanged. In no case were the answers to opinion questions imputed.

Still, there were important variables where missing NSAF responses were imputed in order to provide a complete set of data for certain analyses. For example, the determination of poverty status is crucial, but the income items that had to be obtained to make this determination were often incomplete. In fact, the income item nonresponse could range from 20 to 30 percent in the NSAF; hence the problem could not be ignored. As in other household surveys, NSAF questions about sensitive information such as income, mortgage amounts, health care decisions, etc., encountered significant levels of item nonresponse.

The imputation of missing responses is intended to meet two goals. First, in order to make possible the main goal of the survey, to produce reliable estimates for those living in families with incomes below 200 percent of the federal poverty level, complete data on family income is required. Second, data were imputed to facilitate analyses since complete data are easier to analyze than incomplete data. Without complete data, the analyst is left with three choices for

dealing with missing data, none of which are attractive to most analysts. One can drop variables from the analysis that have missing data, but this runs the risk of omitted variable bias (and could leave the user with very few or no variables to analyze). A second choice is to simply ignore the presence of missing data. By default most statistics packages perform listwise deletion in which all cases which do not have complete data on all variables in the analysis are removed. This procedure runs the risk of bias due to the possibility that the characteristics of persons with nonresponses may differ from those with responses (Oh and Scheuren 1980a and 1980b). A final choice is to model the missing data mechanism explicitly as part of the analysis. Unfortunately, this approach requires a thorough understanding of the process by which respondents do not answer specific survey items, about which most analysts will have little intuition. Furthermore, this last approach is difficult to implement when missing data are not restricted to just a few variables but are instead spread throughout the data matrix (King, et al. 1998).

Organizationally this chapter is divided into six parts. An overview of the hot deck imputation procedure used in the imputations is given in Section 2. The details of the imputations made to do the sample weighting come next (Section 3). The remaining imputations are described in Section 4. Section 5 provides a summary and bridge to the last chapter in this report, which looks at analysis issues, including variance impacts.

## 3.2 Hot Deck Imputation

The approach used to make the imputations for missing responses in the NSAF was "hot deck" imputation. In a hot deck imputation the value reported by a respondent for a particular question is given or donated to a "similar" person who failed to respond to that question. The hot deck approach to imputing missing values is the most common method used to assign values for missing responses in large-scale household surveys. It is the method used for the March Current Population Survey (CPS), the source of the official annual estimates of the poverty population. (See, for example, U.S. Census Bureau, Current Report Series, P-60, No. 198.) For an assessment of the CPS hot deck bias and variance properties, see also Welniak and Coder 1980, and Oh and Scheuren 1980a and 1980b. Recent work on other imputation procedures may hold the interest of some researchers, especially those who may want to redo the imputations (e.g., Little and Rubin 1987). Since all imputed entries are identified, it will be possible to reimpute one or more questions once the full set of public use files has been made available. See Chapter 4 and the general references for more details.

There are many variations of the basic hot deck imputation process. For the NSAF, a hierarchical statistical matching design was used (Coder 1999). The first step in this imputation process is the separation of the sample into two groups, those that provided a valid response to an item and those that did not. Next, a number of matching "keys" are derived based on information available for both respondents and nonrespondents. These matching keys vary according to the amount and detail of information used. One matching key represents the

"highest" or most desirable match and is typically made up of the most detailed information. Another matching key is defined to be the "lowest" or least desirable. Additional matching keys are defined that fall somewhere between these two, and together these keys make up the matching hierarchy.

The matching of respondents and nonrespondents is undertaken based on each matching key. This process begins at the most detailed level and proceeds to less detailed levels until each nonrespondent has been matched to a respondent. The response provided by the donor matching at the best (most detailed) level is assigned or donated to the nonrespondent. For the most part, respondents are chosen from the "pool of donors" without replacement. However, under some circumstances, the same respondent may be used to donate responses to more than one nonrespondent. This can occur if there were no other (unused) donors at the level at which the match occurred. By design, donors are not used more than three times and rarely more than once or twice. As noted already, an imputation "flag" is provided for each variable handled by the imputation system so that imputations can be easily identified.

Sets of matching keys were chosen by examining the outcomes of testing alternative sets of variables making up the keys and by practical concerns. Testing of alternative sets of keys was carried out by selecting a sample of respondents, blanking out their reported information, imputing a new value, and then comparing the imputed value to the value that had been reported. As an example, the most detailed level matching key used for the imputation of wage and salary income from the primary job included variables defining (1) survey site, (2) poverty screener, (3) age, (4) sex, (5) occupation, (6) weeks and hours worked, (7) education, and (8) receipt of means-tested cash assistance. The lowest level matching key for this income source included (1) sex, (2) age, (3) weeks and hours worked, and (4) receipt of means-tested cash benefits.

### 3.3 Imputation Procedures Needed for Weighting

Some items with missing values had to be imputed in order to carry out the weighting process that used these variables. The variables included were: race/ethnicity, tenure (home ownership), additional residential telephones in the household, sex, age, and education (for adults only).

The imputed values were used at the end of the weighting process during the final adjustments to outside controls (see Report No. 3). As a result, missing values were imputed only for completed interviews. Nonrespondents were not weighted in this step, so the imputation was not required for these cases. In general, the level of missing data that had to be imputed to carry out the weighting was relatively small.

It might be noted that if the amount of missing data is small and the data are missing at random (i.e., within groups defined for imputation, the missing data have the same distribution as those with complete data), then the bias of the estimates should not be large. There will still be a

negative effect on the variance due to the imputations; but again, if the frequency of an imputation is low, its impact will also be small (see Chapter 4).

The process was to first use other information reported by respondents in the interview to deduce or edit in the correct value for the person with missing data. For example, if everyone else in the household has the same race/ethnicity, then imputing that value for the person with a missing value is very plausible. Another example is imputing when sex is missing. If the spouse of the person has a known sex, then the missing sex is logically imputed as the opposite sex (see the note to Table 3.1 for more examples).

When no other data in the case were deemed appropriate for filling in the missing value, hot deck imputation methods were used. In these cases, all the cases were sorted into cells defined by characteristics of the respondents, using variables that were as highly correlated with the item being imputed as possible. A missing value was imputed by assigning the value reported for another member of the same cell. If sorting the file into cells created homogeneous groups with respect to the value being imputed, then the imputation should be effective at reducing nonresponse bias.

Two of the variables (tenure and additional residential telephones in the household) are reported once at the household level since they are the same for all members of the household. These items were imputed once at the household level, and the imputed value was assigned to each of the members in the household. Both of these variables were imputed using hot deck methods.

The other variables were at the person level. In general, data from other variables related to the missing item and from other household members were used first in the imputation process. If this information was not available or there were no other household members, then hot deck methods were used to impute the missing values.

In addition to developing keys such as those described above, the imputation design required that imputations take place within key population subgroups. These include:

1. The most knowledgeable adult (MKA) in Option A households (those with children under 18 years of age
2. The spouse or partner of the Option A MKA
3. The MKA in Option B households (those without children)
4. The spouse or partner of the Option B MKA
5. Other household members
6. Focal or sample children.

The order of imputation was governed mainly by the questionnaire sequence. This made sense since the applicable universe of persons eligible for each question often depends on answers to previous questions. For the most part, therefore, imputations were made question by question beginning at the first question within a sequence and proceeding to the last. When a missing

response was imputed, the imputed information was then used in the development of matching keys for subsequent questions, because the responses are often highly correlated. Except for rare situations, persons with imputed values are prohibited from being used as a donor from which imputations are made.

Table 3.1 at the end of this chapter provides the frequency of each imputation for each variable used in the weighting. Shown are both hot deck and logical or edit imputations. Most of these logical imputes are very minor. Just seven persons were assigned gender based on their name and the gender of the other spouse. For race, the editing was more extensive. By design, the Round One questionnaire did not ask about the race or hispanicity of focal children if the respondent and the respondent's spouse or unmarried partner were the biological parents of the focal children. In families where both parents were of the same race and/or hispanicity, we simply assigned the parents' race and ethnicity to the children. In families with different combinations of race and ethnicity, we used regression models based on data from the 1990 Census, to determine which race and ethnicity to assign to the children. This type of editing occurred prior to imputation of race and ethnicity. In all, 1,009 observations were edited in this manner. In the 1999 NSAF, we changed the race and hispanicity questions to ask directly about race and ethnicity for focal children (see Report No. 12 in this series for details).

## 3.4 Other Hot Deck Imputations

There were four additional areas of the NSAF questionnaire where hot deck imputations were carried out; all employed customized SAS macros. Each of these is outlined below, along with how often they had missing information:

Imputation for Income and Employment—Sections I and J of the NSAF questionnaire contain questions covering the current employment and earnings situation. Also included are characteristics of employer-provided health insurance coverage, employment and earnings for the previous calendar year, and source and amounts of income received from "unearned sources," also for the previous calendar year. All of the missing responses in this area have been imputed. The level of nonresponse in these two sections varies considerably depending on the question. For example, for some items, such as current employment status, only 0.2 percent of the responses were missing. For other questions, such as earnings from the primary wage and salary job, nearly 30 percent of the responses for the amount received last year were missing. This nonresponse rate for wage and salary income is nearly the same as that encountered on the CPS. The next highest item nonresponse rate in this section was the amount of income received from Aid to Families with Dependent Children (AFDC).

Imputation for Health Insurance and Health Care Utilization—Imputation for missing responses covered a selected group of questions in Sections B, E and F. These questions included a summary current health insurance source variable derived from

questions E1 – E36 and E37 – E43 as well as health access and use variables from questions F1 – F18, F21, F24, and F27. Imputations were made first for adults and then for children. Variables used to stratify nonrespondents and donors in the hot deck included age, race, sex, poverty status, family size, weeks and hours worked, marital status, health and disability status, geographic location, receipt of means-tested cash benefits, etc. The nonresponse rate for current heath insurance coverage was less than 1 percent for adults.

Imputation for Housing and Economic Hardship—Missing responses were imputed for a small number of questions in Section M. These questions included M6 and M10 – M11. About 12 percent of the responses to question M6 (the amount of monthly mortgage or rent) were missing and imputed. For item M10 (the economic hardship question) about 6 percent of the responses were missing. Most of these nonresponses were generated when the universe for question M10 was expanded after interviewing had been completed for some households. Variables used in the imputation of missing items in this section included age, race, sex, marital status, housing tenure, geographic location, family size, receipt of means-tested cash benefits, size of rent or mortgage payment, postponement of medical expenses, etc.

Imputation of foreign born status (and related items)—Coming as it does at the end of the interview, the foreign born questions occasionally went unanswered because the respondent grew tired or the call had to be interrupted and was not resumed later. Since the race questions also had the same problem, we examined the use of the donors selected for the race imputations but in the end employed a stronger approach based on more information. In all, for 1.3 percent of individuals country of birth and related items were not ascertained. The variables we used were age, ethnicity and race, poverty status, site, metropolitan or nonmetropolitan residence, size, and type of household (husband-wife, single parent with children, other).

As mentioned before, the hot deck imputation system matches nonrespondents and respondents based on a known set of characteristics and assigns the respondent's answer to the nonrespondent. The keys used in the matching were character strings developed by concatenating the various variable values used to define the key. The key could be derived from a maximum of eight different variables. The hierarchical dimension of the system permits matching of respondents and nonrespondents based on multiple keys. These keys differ in the number of variables used in their construction. A maximum of five different keys could be used (of up to eight variables each). This means that a maximum of five attempts were made to find a donor with each succeeding attempt employing a less detailed key.

The imputation of missing responses required a large number of steps. These steps are most often defined by a SAS program that either performs edits prior to imputation, prepares data needed in development of imputation-matching variables, generates the actual imputations, or prepares a "final" file containing the post-imputation variables and flags that can be matched to the master file.

Table 3.2 (at the end of this chapter) lists all the portions of the NSAF questionnaire for which imputations were conducted. Detailed discussions and descriptions of each step are found in Coder (1999). Below, to give a flavor of the extensive efforts involved, the imputation of the employment and income portion of the survey is covered. This included the income screener question plus all of Sections I and J of the extended interview.

### 3.4.1 Household Screener — Income Question

This imputation was done first before anything else because of its central importance in later imputations. Since the income screener question was one of the main variables used to impute later detailed income amounts, it was addressed first. Of the 44,461 households, 2,011 were missing the response to the income screener question. All of these missing responses were imputed.

The actual imputation was divided into two phases. In the first phase, the value from the extended interview of the family income (FAMINC) variable was used to make an assignment. In cases where the value of FAMINC was greater than the income screener threshold for households of the same size, the screener value was assigned as "above 200." In these cases one could be sure that the household income would exceed the 200 percent of poverty threshold used in the screener. For cases where the value of "FAMINC" was greater than zero but less than 75 percent of the income screener, the screener value was assigned as "below 200." In the second phase of the imputation, those cases that remained as nonresponses after the first phase were imputed using the standard hot deck methods employed throughout this imputation system.

### 3.4.2 Section I — Employment Questions

Virtually all questions in Section I have been imputed if missing. Not imputed were "auxiliary" type questions that were only asked as a fallback when the response to a previous question was "Don't know." An example of this would be the questions in the I15 – I16 sequence where the question on "the number of persons employed" at the respondent's place of work is followed by a question to try to obtain a response indicating whether or not the number is greater than or less than 50. There are other items that were not imputed as well, but very few.

The number of items contained in this section is extremely large. For that reason, the imputation process was broken down into segments. Each of these segments often contains a large number of steps defined by SAS programs that perform the basic functions noted earlier (see Coder 1999 for details).

The imputations for Section I were separated into six separate tasks. These are shown below as follows:

1. Impute part 1 of Section I Current Employment Questions (excluding earnings)
2. Impute part 2 of Section I Current Employment Questions (excluding earnings)
3. Impute Weeks Worked Last Year of Section I Questions
4. Impute part 3 of Section I Work Experience Last Year (excluding earnings)
5. Impute Section I Earnings Amounts Questions (current and last year)
6. Impute Section I Other Person Work Status and Earnings Questions.

In the first four tasks involving Section I, the imputation of all items **other than earnings** amounts is undertaken. The imputation of missing earnings amounts takes place in task five for all sample persons and in task six for nonsample persons. The task of imputing missing earnings amounts was consolidated into a separate process so that the earnings amounts for both current and last year's employers and businesses could be imputed together.

The separation into tasks was partly driven by the subject matter and partly by a need to limit the computing time required to complete the imputation process. Since most of the programs making the imputations require considerable amounts of time, it was reasonable to build, test, and implement them in pieces in order to minimize the time needed to rerun a program should an error be encountered or a revision required. This was often the case, and this proved to be a good strategy. (Incidentally, a modular strategy was also followed for other sections of the questionnaire as well.)

The imputation of these items took place in sequence beginning with the current employment question which begins Section I. The "skip patterns" and universe definitions were used to determine which questions were applicable. As noted above, the imputation of earnings questions was undertaken following completion of the imputations for all other Section I questions.

The universe and related skip patterns (see Report No. 12) were not programmed into the computer version of the questionnaire with 100 percent accuracy. The reason for this was that the survey evolved in its early weeks since the NSAF was a first-time effort. In general, while there were only a small number of inconsistencies on the data file, they caused a lot of problems in the development of the imputation programs.

The level of missing responses varied greatly in Section I. The number of nonresponses to the "currently employed at a job or business" question was extremely low. Much higher levels of

nonresponse were found for establishment size and many of the health insurance – related questions.

The imputations for Section I took place at a point in the file development when age and educational attainment variables had not been imputed. If these items were missing but required in the development of matching keys, the missing values were replaced by "median" values in order to permit the imputations to go forward. This practice was in lieu of keeping the nonresponses as a separate category in the key. Since the levels of nonresponse were low for these items, the pool of donors defined by nonresponse would have been too small to achieve any reasonable matches. In the second round, the imputation of missing age, sex, race, education, and other key socioeconomic variables will take place soon after the data are received in order to avoid this problem.

### 3.4.3 Section J — Imputation of Income Recipiency and Amounts

Section J consists mainly of a series of questions designed to determine who in the family received income from a large number of transfer programs and other unearned sources. For each item, the annual amount received for each person was requested, and in many cases the number of months or weeks during which the income was received. The income amounts could be reported as either the monthly amount received or the total annual amount received last year. The exceptions were emergency assistance and financial assistance from friends and relatives. The separate payment amounts (up to four amounts could be recorded) were combined into a single value. Imputation of missing amounts for these two income sources assigned a single annual amount.

Specialized treatment was required for the income sources reported in question J18, that is, income from some other source not covered by the specific questions. A write-in entry was recorded to describe the source and the total annual amount was recorded in question J64. All of the write-in entries for these cases were examined and a numerical income source code was assigned.

Imputation for missing responses in Section J required four major tasks. The first task was to perform a consistency edit on the responses and to put them in a consistent format before sending the data to the imputation programs. The second task was the imputation of missing income recipiency for the reference year. There were a small number of nonresponses both at the family level and person level involving income recipiency. It is not clear exactly how the recipiency responses were formed given that the first question determined income receipt at the family level and the second question established which persons in the family received income. In cases where the family response was missing, any person-level responses indicating "not in universe" were changed to nonresponses and later imputed.

The third task was imputation of missing amounts, periodicity of payment, and number of months received for each source. These imputations took place sequentially but not in the order on the questionnaire. Instead, the order of imputation was chosen to provide maximum consistency between amounts from different income sources. For example, interest and dividend amounts and other non-means-tested income amounts were imputed before means-tested income sources were imputed so that these amounts would be available for use in developing variables to be used in the matching keys.

The final task was the categorization of the "any other income" write-in entries and the imputation of any missing responses for up to four other income sources. A computer coding system was devised to assign codes based on key words. For those cases that could not be assigned a numeric code in the computer-based examination, a manual review of each case was undertaken to assign a code. Imputation for missing responses was made by assigning the median value reported by all respondents. In addition to these edits on the income items themselves, files are merged to retrieve information needed to develop the hot deck matching keys. These include demographic information, household-based information (such as site and income screener), weeks worked and earnings amounts from the post-imputation file generated earlier, etc.

### 3.4.4 Other Important Considerations

The sequential nature of the imputation process is critical since the imputed response to one question affects the flow of the questionnaire from that point on. In this sequential process, the data file is updated following the imputation of each item. That permits the imputed values to be used in determining the proper universe for subsequent questions and to be included in development of the matching keys for any following questions.

As a general convention, imputations were carried out separately for different classifications of sample persons. This approach permitted the use of slightly different sets of matching keys for each group and provided manageable run times on the computer. While different sets of matching keys can be used, as a practical point, attempts were made to use the same sets initially. Note that in some situations the set of matching keys for a particular variable may need to be altered because the specified set does not find a match at any level for one or more nonrespondents.

Editing of the questions was needed prior to imputation; particularly important was changing an item so that it had a "not in universe" code, if it was determined that the person was not in universe in the consistency checking necessary prior to imputation. For example, since the responses to the questions in the current employment (part 1) portion of Section I determine the path followed for questions covering work experience and earnings last year, it was necessary to reassess the universe for the weeks worked questions following imputation. The imputations may have altered the path into this series of questions or resulted in the exclusion or inclusion of

persons into the section altogether. This reassessment was also deemed necessary because the questionnaire paths indicated were, in a small number of cases, inconsistent with the questionnaire outcome.

The first part of this reassessment examined the entries in the "earnings last year" questions and established a simple "work status last year" variable based on the presence of earnings last year. If any earnings amounts were something other than zero or "not in universe," then a worker status was assigned. This assured that the imputation, if necessary, would result in a nonzero weeks imputed value.

The second part of this reassessment edit was a universe and consistency edit for questions I49a–I50. For persons who were determined to be out of the universe for the weeks worked series, all questions related to work and earnings were set to the "not in universe" value. For those determined to be in the universe for the weeks worked questions, a consistency edit was completed for the weeks worked sequence.

The consistency edit was followed by the imputation of missing responses. As mentioned above, a worker status variable was generated for all persons based on responses to the earnings questions. The imputation process included two steps. The first step involved only those persons having a work status assigned as "worker." All workers with a missing response to items in the sequence were assured of being imputed as nonzero weeks worked. The second step imputed responses to those persons whose worker status could not be determined in the examination of the earnings amounts fields. For these persons, the imputation of weeks worked could result in a zero value. Since the response to the weeks worked question is accompanied by a question on "units" (either weeks or months), both of these questions were imputed jointly.

## 3.5 Summary

In this chapter we have illustrated the care with which the imputations were conducted. Our aim was to illustrate, but not exhaustively document, the imputations and edits made. The references cited in this report should be consulted if more detail is desired, notably the public use files where a variable-by-variable approach is taken.

Despite limitations of scope, several elements stand out and may be worth mentioning in this brief summary:

First, the imputation process is a well-known and commonly used one. It has many similarities, for example, to the approach used in the CPS.

Second, the process was applied over and over again to reflect the different nature of the "missingness" encountered.

Third, the imputations made the greatest use possible of interrelationships—whether logical or stochastic—when these existed between variables.

Fourth, while not always achievable, every effort was made so that assignments would be consistent with the existing information and with each other.

Fifth, the process was sequential and sometimes iterative, as more was learned about the survey.

As we mentioned in the introduction to this chapter, the imputations made are believed to have been reasonably successful in dealing with the biases caused by item nonresponse. This does not mean, as we will see in the next chapter, that the imputations made can be treated as if they were real data obtained from respondents. The imputations made by the hot deck approach increase the sample variance of any analyses done, and they can also, if not employed properly, increase the potential for misclassification errors. These and other limitations are covered in the next chapter.

**Table 3.1 — Imputed or Edited Data when Missing for Variables Used in Weighting (In Percent)**

| Variable | Imputed (Stochastic) | Edited (Deterministic) |
|---|---|---|
| Own or rent | 1.5 | None |
| Other phones | 0.91 | None |
| Business phones | 1.1 | None |
| Gender | 0.05 | 0.01 |
| Age | 0.50 | 0.03 |
| Race/Ethnicity | 2.6 | 18 |
| Specific Race | 3.4 | 22 |
| Education | 1.4 | 0.05 |

Note: In this table, we distinguish between two ways in which missing data are filled in - either stochastically or deterministically. Imputed cases are those which were imputed either through a hot deck imputation or, in the case of race/ethnicity and specific race, were assigned through the use of allocation algorithms (which required the use of random number generators as part of the allocation process). Examples of deterministic edits include using the child's current grade in school to assign age (the modal age for a given grade was assigned) and for race/ethnicity and specific race, when these variables were filled in based on the values of other members in the household (and no randomization was required).

**Table 3.2 — Number of Additional Items Imputed by Section — 1997 NSAF**

| Section | Items Imputed |
|---|---|
| **Income Screener** | |
| Income Screener Question | 1 |
| **Sections I and J** | |
| Part 1 of Section I Current Employment Questions (excluding earnings) | 29 |
| Part 2 of Section I Current Employment Questions (excluding earnings) | 12 |
| Weeks Worked Last Year of Section I Questions | 5 |
| Part 3 of Section I Work Experience Last Year Variables (excluding earnings) | 12 |
| Section I Earnings Amounts Questions (current and last year) | 8 |

| | |
|---|---|
| Section I Other Person Work Status and Earnings Questions | 2 |
| Other Income Section J Questions | 54 |
| **Sections E and F** | |
| Current Health Insurance Coverage Summary Variable (based on E1–E36) | 36 |
| Health Insurance Questions (E37–E43) | 7 |
| Unmet Medical Need Questions (F19–F29) | 11 |
| Health Care Utilization Questions (F1–F17) | 17 |
| **Section M** | |
| Question M6 (amount of mortgage or rent) | 1 |
| Questions M10–M11 | 2 |

Note: For each variable imputed in each section of the questionnaire a flag was inserted on the computer file. Full details on these will be found in the public use file documentation. To have provided all of theses counts here was considered but for reasons of space they have been omitted. We have, however, provided information on the items included in the first two public use files (documented Reports No. 11 and 13, see Tables 4.5 through 4.17 at the end of the next chapter.

# Chapter 4
# Analytic Concerns Related to Imputation

## 4.1 Introduction

In general, as we have seen from the calculations in Report No. 4 of this Methodology Series, NSAF analyses can be affected by the fact that the data are based on just a sample, albeit a large one. Moreover, the survey is subject to the problems of undercoverage, unit nonresponse, item nonresponse, and measurement or response error.

In a sense, the whole methodology series has, as its goal, the documentation of how each of these and related concerns have been addressed and how they can be managed by end-users of the survey. For example:

> The sample design and variance estimation reports (Report Nos. 2 and 4) deal mainly with the sampling component of the NSAF total survey error.
>
> The problem of undercoverage is dealt with in two estimation reports (Report Nos. 3 and 14)—the latter of which is forthcoming.
>
> Unit or noninterview nonresponse is also the subject of two reports (Report Nos. 7 and 8)—one on the response rates achieved (No. 8) and the other on the extent to which there were residual biases (No. 7).
>
> Measurement issues surface in many places, notably in each of the public use file codebooks (e.g., Report No. 11) and in the discussion of the 1997 and 1999 NSAF questionnaires (Report No. 12).

The material here is intended to link up with the documentation being prepared for the public use files from the survey. It also can be seen to parallel best practice in the survey field as represented by the work done by Tom Jabine (and others) on what have come to be called survey "Quality Profiles" (Jabine 1994).

As mentioned earlier, Report No. 11 documents the first of these files and in many ways is a prototype of the others now underway. All the public use files have codebooks that take a variable-by-variable, approach noting inconsistencies and editing issues. If the item has been imputed, counts are provided and a flag is available for researchers to reimpute the item, should they wish to do so. When there are known to be other measurement concerns, these are also covered variable-by-variable.

Organizationally, the present chapter is divided into five parts. This short introduction (Section 1) begins our treatment. A discussion of general statistical considerations follows in Section 2. Next is a rough

guide to the variance impact of the hot deck imputations on the NSAF variables (Section 3), along with a fully worked example of what it would take to approximate the imputation variance (Section 4). Section 5 concludes the chapter and the report. Selected references to the extensive technical literature on editing and imputation are cited throughout.

## 4.2 General Statistical Considerations

It is possible to distinguish three types of measurement issues in the 1997 NSAF. Each can have different implications for end users and hence deserves separate discussions:

- Misreporting and undetected inconsistencies across survey items
- Missing item information that was not imputed
- Missing item information that was imputed.

Each of these is taken up briefly below and some speculations are given that may be of use to researchers in helping them with their analyses.

### 4.2.1 Misreporting and Undetected Inconsistencies

As noted in Chapter 2, both Westat and the Urban Institute implemented careful quality control procedures at every stage of the survey process. Among other things, extensive computer checks were conducted to detect and correct inconsistencies or illogical patterns in the data. This does not mean that all such erroneous patterns were uncovered. In fact, it is important to note that complete consistency between all questions and the CATI specifications (see Report No. 12) may not have been achieved in the post-imputation data set.

Standard graphical and other outlier analysis efforts will be needed by researchers to protect against the possibility of inconsistencies since we could not anticipate all the ways the NSAF data would be used. We believe, however, that any outliers found will be infrequent. We would greatly appreciate being informed of any such inconsistencies researchers may uncover, so they can be brought to the attention of others and, depending on their nature, new data sets made available.

Misreporting errors that do not lead to inconsistencies may remain as well and researchers should be on the lookout for these, too. One example here is the evident heaping of age for NSAF-sampled adults (especially at 30, 40, etc.). This problem arose because we asked for age directly rather than date of birth. The problem is judged to be a minor one for most researchers, since the data have been reweighted to align the survey totals with outside population controls. In Report No. 13 in the

methodology series, which documents the second public use file, there will be more about this particular error and the impact it has. See also Report No 11.

Some misreporting, unlike that for age, required changes to the questionnaire between the 1997 and 1999 NSAF rounds. Particularly important examples of this were the foreign born questions. The questions used in the 1997 NSAF were not believed to have identified all foreign born individuals. This led to a new sequence of questions in the 1999 NSAF. See Report No. 12 for more specifics here, including some of the fairly successful (if partial) strategies for handling this problem in the 1997 survey.

### 4.2.2 Missing Item Information That Was Not Imputed

As has been said earlier, for most NSAF questions item nonresponse rates were very low, often less than 1 percent. Basically, unless it was needed to weight the sample, items with very few missing entries were left unchanged. In no case were the answers to opinion questions imputed.

Given that this was our procedure, what should a researcher who wishes to analyze one or more questions with missing information do? As might be expected, there is no one answer we can recommend. We do, however, emphasize the following emphatically: ***Ignoring this issue may lead to inaccurate interpretation of the data.***

Some investigation of the "missings" in the context of a researcher's particular analysis almost certainly should be done. Simple (weighted) tabulations comparing the group with missing information to the rest of the sample may be all that is needed.

For some questions, after a quick look, the decision to do nothing more may be the right one. In the case of opinion questions, for example, the level of nonresponse is very small and it seems possible to restrict inferences just to those expressing a viewpoint. However, it is good practice to consistently report the weighted percentage of an item that was not answered.

By ignoring the missing data (i.e., by restricting your analyses to only complete cases), the implicit assumption is that those with missing information are on average like those responding. Imputation adjusts for differences between those who provide an item and those who do not. The effect of this can be seen by comparing the reported and imputed data. This approach has been taken here for the variables with missing data imputed in the first two public use files. In particular, see Tables 4.5 through 4.17 at the end of this chapter. Clearly these tables and related figures illustrate the point, since the imputed cases are different from those that reported the item to begin with or, at least, they have imputed values that have a different distribution overall.

Techniques, such as those used to weight the NSAF sample (see Report No. 3), may be employed for a particular analysis, with a special reweighting done. Such reweighting would not generally be multipurpose (Little 1986; Scheuren 1986). Another specially focused option is to model the "missingness" explicitly as part of the analysis (e.g., see Little and Rubin 1987). Such approaches are

challenging to implement, however, and would only be recommended in cases where there was a concern that a simpler approach, say, a form of multi-purpose imputation (as we have done), would not be adequate. For researchers wishing to do their own imputations the books by Rubin (1987) and Schafer (1997) are recommended.

Imputation is an area of very active statistical interest and in fact there is an electronic mail listserv, called IMPUTE, devoted entirely to the subject. Researchers wishing to do their own imputing or reimputing may want to "listen in" or even post their own NSAF questions directly.

Many of the best people working in this area of surveys appear to be participating regularly. To subscribe, the following procedure should be followed:

Send an email to listserv@listserv.nodak.edu and in the body type SUBSCRIBE IMPUTE [your name]. A sample email follows:

From: johndoe@email.com
To: LISTSERV@listserv.nodak.edu
Subject: SUBSCRIBE IMPUTE

SUBSCRIBE IMPUTE John Doe

One concern to call to the attention of those wishing to do their own imputation is that the survey weights should be considered in the imputation. In the NSAF hot decking, consideration was given to the weights by using the major stratification variables in defining imputation cells–variables like site and poverty status. There is literature on employing a weighted hot deck which may also be worth consulting (see, for example, Cox and Folsom 1981 and Williams and Folsom 1981).

The use of an unweighted hot deck our preferred option, is predicated on the notion that the weights within the hot deck cells being used are fairly similar among the donors, or that the cell definitions were very predictive. To safeguard against the failure of this model, the cells chosen ought to include the variables used as strata in the original selections (e.g., site and poverty status) or the post-strata used in the estimation (e.g., race, gender, and age).

This was the basic approach in the NSAF, and it should offer sufficient protection against model failures in the hot deck matching, provided the amount of collapsing is not extreme. As a diagnostic, researchers might want to indicate at what level of matching the hot deck assignment took place. In our imputations for the 1997 survey, we did not do this but are considering employing this level of auditing in the second round of the NSAF. It virtually never happened, however, that we made imputations outside the original design strata defined by site and poverty status.

### 4.2.3 Missing Item Information That Was Imputed

The hot deck imputations for missing responses in the NSAF are an attempt to adjust for biases that may result because the characteristics of persons with nonresponses tend to differ from those of respondents. In this endeavor, we believe the imputations are at least partially successful. They also unquestionably make the NSAF data much easier to use. Furthermore, a common set of imputations can make comparability across separate NSAF research efforts easier to achieve. The reason for this is that analysts can all employ the same imputed values. If researchers elect to do their own imputation they may still want to report both their own results and what NSAF, as provided, would produce.

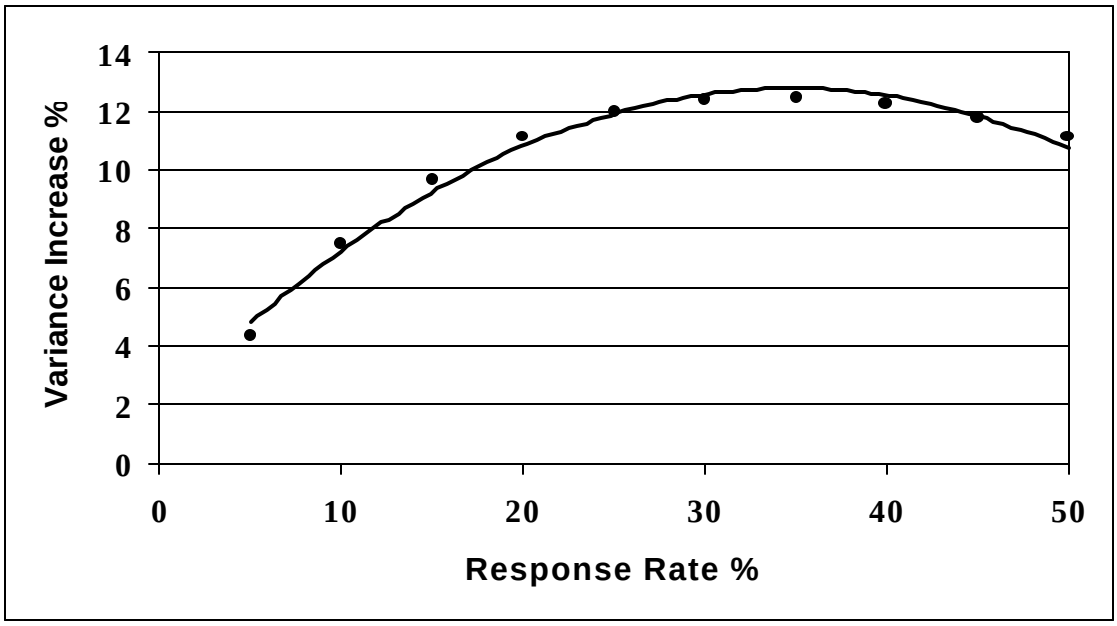
This ease of use can be a problem because it may lull some analysts into the belief that the imputed values can be treated just as if they were real. This is far from true. Not only are there likely to be residual biases in the nonresponse, but the sampling variance is affected as well. What to do in the NSAF about the variance impact of imputation is the topic of the next two Sections (4.3 and 4.4). Concerns about bias impacts, which are potentially more serious are covered briefly in the concluding Section (4.5).

## 4.3 Upper Limit Imputation Variance Impacts

There are two sources of increased sampling error that arise due to hot deck imputation. First, the procedure is subject to Monte Carlo error because within each hot deck cell, the donors are selected at random. Second, there is the fact that we are not adjusting for the smaller (respondent) sample size in the variance calculations. Each of these concerns is taken up separately below and subsequently their combined influence is illustrated.

> Monte Carlo Error.—If another donor had been selected, another imputation would have been made and presumably a different value obtained. This source of variance increase is generally modest, provided the imputations are done such that the same donor is seldom reused. In some cases (Hansen, Hurwitz, and Madow 1953), theoretical results are available which bound this source of error. In fact, under the assumption of simple random sampling and the selection of donors without replacement, the maximum value of this source of variance increase is 12.5 percent, which occurs when one-third of the cases have missing values. Figure 4.1 plots the percent increase in the variance as a result of this factor. The figure only shows values up to a response rate of 50 percent. Above this level the assumption of sampling without replacement cannot hold. See, however, Oh and Scheuren (1983) for extensions.

**Figure 4.1 — Variance Increase From Hot Deck Random Assignments Versus Response Rates**




We will use the Hansen, Hurwitz, and Madow (1953) results as one element in constructing a rough upper bound on the variance impact of item imputation. This has been done in Tables 4.1 and 4.2 at the end of this chapter.

Adjusting Sample Size.—One of the most troubling impacts of a hot deck imputation is that it invalidates the assumptions under which the sampling variances are calculated (see Report No. 4). The software used in variance estimation, whether Wesvar or something else, assumes that the imputed values are real and instead of using the actual sample size, say "m," it treats the total sample of size, say "n," as complete. Thus, *ceteris paribus*, in simple random sampling we should be adjusting the calculated sampling variances upward by the factor n/m.

Illustration.—To illustrate, suppose that the assumptions above were true, and furthermore that we had a sample of size of n = 3000 where a third of the cases had to be imputed so that m = 2000. Also suppose that we did the imputation such that the maximum value of the Monte Carlo error was 12.5 percent. This would mean that the variance estimates obtained routinely would have to be increased by the very large factor

$$(1.125)(3000/2000) = 1.67.$$

Confidence intervals would be similarly increased but by the square root of this factor, or by about 30 percent. Clearly, such sizeable effects cannot be ignored. (Fortunately, this extreme does not occur in the NSAF.)

Of course, if item nonresponse is very small, neither of these concerns make much difference. To see this, suppose that we look at the items imputed on the first two public use files (Reports Nos. 11 and 13). This has been done at the end of this chapter. Under the assumption that we can ignore the complex nature of the NSAF sample design and of the hot deck imputation procedures and treat both as if the data sets were simple random selections, the maximum variance increases are modest—no more than 4 percent. The confidence intervals are lengthened no more than 2 percent. However, for some variables, like earnings, this would not be the case (see Section 4.4).

Tables 4.1 and 4.2 may be useful as upper limits on the imputation variance contribution, but it can be argued that they are potentially quite conservative. To see this, suppose the model used in the hot deck imputation predicted the missing values exactly. Then, in the sense we have used it here, there would be no "imputation error" at all and hence no increase in variance. Now, in general, suppose we had to impute Y as a function of a set of variables $\underline{X}$; that is,–

$$Y = f(\underline{X}) + \text{error}.$$

In the present context, the function "f" could be defined by our hot deck cell selections. But it could also be a regression function or something similar. Now if $f(\underline{X})$ predicted Y without error, then the survey might not have needed to ask Y. Since such a redundancy would have cost money and contributed nothing, it would undoubtedly have been weeded out beforehand.

While it is unreasonable to suppose that perfect predictions happen in practice, it is equally unreasonable to suppose that the carefully chosen imputations done have not captured at least some of the underlying structure of the phenomenon. Some predictive power undoubtedly exists in the set of $\underline{X}$'s, therefore, and our upper limit formula has not allowed for this.

Subject matter knowledge and a detailed analysis of the imputations themselves might be needed to refine these limits. Provided the amount of missing information is small, though, these refinements may not be necessary. On the other hand, when the amount of "missingness" is large, another approach altogether — a version of multiple imputation (Rubin 1987) — may be worth examining. This is covered in the next section.

## 4.4 Multiple Imputation of Earnings

In this section we examine one of the key survey variables that required a considerable amount of imputation. This variable was last year's earnings. Of the 75,525 adult respondents in the NSAF, there were 61,021 who worked last year and were, therefore, in the universe applicable to the earnings questions. A total of 17,949 were imputed earnings amounts in the imputation system. This means that the rate of item nonresponse was

$$(17,949)/(61,021) = 29.4\%.$$

This made earnings the most frequently imputed item in the NSAF (by far); hence, it is a natural candidate for further examination.

To set the stage for this examination, we first describe the basic (quite complicated) imputation procedure employed, then go on to analyze what happens when we reimpute—at random—nine times. Finally, we calculate the conventional (Wesvar) sampling error and speculate about how much it should be adjusted, given the imputation error introduced.

### 4.4.1 Original Imputations Designed for Earnings

A group of questions in Section I of the NSAF questionnaire is designed to collect the annual earnings amount received "last year" for each respondent. There are two questions covering wage and salary income, one question covering self-employment income, and one question covering earnings from all other jobs or businesses. The wage and salary questions and the self-employment question are designed to obtain the earnings from the main job or business and are mutually exclusive (each of the separate wage and salary questions applies to somewhat different universes depending on the respondent's employment situation at the time of the survey). Only one of these three questions is applicable to an individual who had worked last year. All respondents working last year are also asked the question covering earnings from all other jobs or businesses.

Attempts are made to bound a respondent's annual earnings amount when they were unable or unwilling to provide a point estimate. In these cases, a series of questions were asked to place the earnings amount: (1) under the poverty level for the family, (2) between the poverty level and 2 times the poverty level, (3) between 2 times and 4 times the poverty level, or (4) more than 4 times the poverty level.

The imputation process then assigns values to missing responses to the appropriate earnings questions. There are two basic stages in the imputation process for earnings. Stage 1 imputation applies to persons providing the poverty bounding information. Stage 2 applies to persons who did not provide the poverty bounding information or who failed to obtain a matching donor in stage 1. Within each stage respondents are separated into the basic NSAF respondent groups: (1) MKA, (2) spouse/partner of MKA, (3) Option B or adult respondents without children, and (4) spouse/partner of Option B respondent. The imputation for each group takes place independently.

The hierarchical statistical matching method is used to match nonrespondents and donors. Once nonrespondents and donors within a matching cell have been identified, a "starting donor" within the list of donors is selected. Matching of nonrespondents begins with this starting donor and proceeds through the list of donors as needed. If the number of nonrespondents exceeds the number of donors, then multiple use of donors is permitted up to a maximum of three uses. Multiple use of donors was not common but did occur.

The starting donor within the matching cell was selected "at random." A random number generator provided in the SAS system was used as the basis for making the selection. The generator provided a value between zero and one. This proportion was applied to the number of donors to yield the position in the list of donors where the first donor would be taken.

### 4.4.2 Reimputation of Earnings

In this effort to estimate imputation variance, nonrespondents have been imputed a total of nine times. The only difference for each pass of the imputation system was the "seed" used to obtain the random number employed to select the starting donor. The same seed was kept throughout any given imputation iteration.

The output of the imputation was summarized in a single data file. Each observation on the file includes the iteration number. As already noted, the total universe of MKA/Option B/spouse/partners contained 75,525 respondents. Of these, there were 61,021 who worked last year and were, therefore, in the universe applicable to the earnings questions. A total of 17,949 were imputed earnings amounts in the imputation system. The SAS output data file, therefore, contained 679,725 observations (9 x 75,525).

Each observation on the data file contained the person identifier, the total earnings amount last year, a "flag" indicating the imputation status, and the number of the imputation iteration. The total earnings amount was created by adding together the appropriate earnings amount from the main job or business with the amount for all other jobs or businesses. The imputation flag reflects imputation to either one or both of these amounts. At the end of this chapter, Tables 4.3 and 4.4 summarize what was learned. In those Tables we have produced distributions by earnings size class from each imputation, plus an overall average—Table 4.3 provides the data on the percentage by earnings class and Table 4.4 shows average earnings within the classes chosen.

### 4.4.3 Wesvar Sampling Variance Compared to Total Variance

The variance of the tabulated earnings distribution was also calculated using the Wesvar methods of Report No. 4. The sampling standards errors are also shown in Tables 4.3 and 4.4, where we have combined them with the imputation error to produce a (DEFT) adjustment factor that displays the degree to which we need to lengthen the confidence intervals being calculated over those that are available from just using Wesvar.

The first and last classes in the table (which are open-ended) show some variance imputation effects with DEFT values that are not close to one. The other classes show modest or small effects, on the order of 5 percent or less in many instances (for percentages) and on the order of 10 percent (for amounts). By their nature, these reimputations may not have captured all of the Monte Carlo error. The reason for this is that at the most detailed level of application there may have been only a single donor. This would mean that for each iteration of the reimputation we would always get the same

donor—hence understating the true imputation variance. Exactly how important these limitations are we cannot say. To deal with these concerns in the future we have added a counter to the imputation programs that will be used in the 1999 NSAF, and the size of imputation cells will be recorded (this would allow adjustments of the sort in Oh and Scheuren 1980a to be made.)

In general, there is some controversy around how to interpret multiple imputation results, even within a full Bayesian framework (e.g., Schafer 1997). We plan to investigate multiple imputation further in the 1999 NSAF. Other approaches such as imputing within each variance replicate separately could be employed with the NSAF. These would have required 61 imputations for each missing value (see Shao, Chen & Chen 1998). Even so, for some items like earnings it may be worth the trouble.

## 4.5 Overall Conclusions

Issues around the quality of NSAF measurements have been covered at length in this report. Chapter 2 has done this for editing and Chapter 3 for imputation. In the present chapter we have attempted to provide a "starter set" for researchers wishing to understand how NSAF measurement concerns related to imputation might affect their analyses.

To conclude it might be worthwhile to reiterate certain points and to emphasize that the NSAF is not unique in its concerns about measurement. Indeed, in many respects the NSAF is similar to other large-scale household surveys—as we will elaborate on further in Report No. 15, where the NSAF is compared directly with the CPS and other major national samples:

- As is evident, the existence of missing and misreported items in the NSAF made for much extra work.

- Both the NSAF data editing and data imputation done should greatly ease the analysis burden of researchers.

- For most imputations the variable impacts of the hot deck imputations used are quite modest and lengthen standard confidence intervals only slightly. Outlier detection should be built into the analyses done. Nonetheless, anomalies are expected to be rare.

- For many items no imputations were done. The "missingness" rates for these items generally were very low. While some checking is advisable, research concerns should be minimal.

- Even for the items for which imputations were done, the "missingness" rates were small. Hence, while again some checking is advisable, research concerns should be minimal.

- There is some evidence (see Table 4.5 – 4.17) that the imputations may have at least partially reduced nonresponse bias.

- In a few cases, earnings is the most important example, the variance impact of imputation can be relatively large.

- If researchers have concerns about bias or variance impacts they may consider redoing the NSAF imputations. Every effort has been made to facilitate this on the public use files.

This said, usually bias concerns dominate in a large survey like the NSAF or CPS. Because biases are so hard to measure, however, we have concentrated on variance impacts, relying on researchers general knowledge to guide them otherwise. It would have been better to have had a combined treatment of both variance and bias (Oh and Scheuren 1980a and 1980b), but such opportunities rarely happen. We did do some bias estimation in the NSAF but this was for the unit nonresponse problem—see Report No. 7. Comparisons between the NSAF and other national surveys will be the topic of Report No. 15 in this methodology series, which will provide an in-depth examination of relative bias issues. The examinations carried out so far (e.g., Report No. 1) reflect very favorably on the NSAF.

**Table 4.1 — Focal Child Public Use Variables: Upper Limit Factors for Imputed Categories**

| Variable Acronym | Number Responding (1) | Response Rate (2) | Loss (3) | DEFF (4) | DEFT (5) |
|---|---|---|---|---|---|
| AGE | 33648 | 99.84% | 0.0408% | 1.0020 | 1.0010 |
| SEX | 33666 | 99.89% | 0.0274% | 1.0014 | 1.0007 |
| UBETH | 32659 | 96.90% | 0.7742% | 1.0400 | 1.0198 |
| UBRACE | 32659 | 96.90% | 0.7742% | 1.0400 | 1.0198 |
| BDISBL | 33559 | 99.57% | 0.1068% | 1.0054 | 1.0027 |
| BHLTHN | 33672 | 99.91% | 0.0230% | 1.0012 | 1.0006 |
| BHLTHP | 33639 | 99.81% | 0.0475% | 1.0024 | 1.0012 |
| MOWNRENT | 32986 | 97.87% | 0.5318% | 1.0272 | 1.0135 |

Note: This table illustrates, for the focal child public use file, what the variance impact is of hot deck imputations. Descriptions of each variable and column follow.

The variable acronyms stand for focal child age (AGE), gender of focal child (SEX), ethnicity of child (UBETH), race of child (UBRACE), has health condition that limits activity (BDISBL), current health status of child (BHLTHN), child's current health compared to 12 months ago (BHLTHP), and own or rent (MOWNRENT).

**Column (1)**: The total number of children with a reported value (out of a total of 33,703)

**Column (2)**: Column (1) divided by the total number of focal children on the file (33,703). Using weighted response rates might be better, in general, but this makes little difference.

**Column (3)**: The formula used to assess the "loss" in efficiency from the hot deck random assignments is given as:

[Column (2)][1 - Column (2)] / [1+ Column (2)]$^2$

This result is from Hansen, M.H., Hurwitz, W.N. and Madow, W.G. *Sample Survey Methods and Theory*, Vol. 2. 139 – 141; in the reference the given formula is $(n_1/n)(1 - (n_1/n)) / (1+n_1/n)^2$, where $n_1$ = number responding and n = total number. We replaced $n_1/n$ with Column (2) since the quantities are the same.

**Column (4)**: Equal to [1/Column (2)][1 + Column (3)]. The design effect (DEFF) is a ratio by which variance should be increased when using the imputed data. The factor 1/[Column (2)] is included to correct for the fact that standard variance calculations software treats the imputed values as real, thus overestimating the sample size. See Report No. 4 for more information on the use of "DEFF" in a more general context.

**Column (5)**: The square root of DEFF.

**Table 4.2 — Most Knowledgeable Adult Public Use Variables: Upper Limit Factors for Imputed Categories**

| Category | Number Responding (1) | Response Rate (2) | Loss (3) | DEFF (4) | DEFT (5) |
|---|---|---|---|---|---|
| AGE | 27209 | 99.31% | 0.1725% | 1.0087 | 1.0043 |
| SEX | 27396 | 99.99% | 0.0018% | 1.0001 | 1.0000 |
| UBETH | 26575 | 97.00% | 0.7508% | 1.0387 | 1.0192 |
| UBRACE | 26575 | 97.00% | 0.7508% | 1.0387 | 1.0192 |
| UBCPSED | 27232 | 99.39% | 0.1515% | 1.0076 | 1.0038 |

Note: This table illustrates, for the most knowledgeable adult public use file, what the variance impact is of hot deck imputations. Descriptions of each variable and column follow.

The first four variable acronyms have the same meaning as in Table 4.1 except they are for the most knowledgeable adult, not the focal child. The fifth variable acronym stands for education level, CPS (UBCPSED).

**Column (1):** The total number of MKAs with a reported value (out of a total of 27,398)

**Column (2):** Column (1) divided by the total number of MKAs on the file, (27,398). Again, using weighted response rates might be better, in general, but this makes little difference.

**Column (3):** The formula used to assess the "loss" in efficiency from the hot deck random assignments is given as

[Column (2)][1 - Column (2)] / [1+ Column (2)]$^2$

This result is from Hansen, M.H., Hurwitz, W.N. and Madow, W.G. *Sample Survey Methods and Theory*, Vol. 2. 139 – 141, in the reference the given formula is $(n_1/n)(1 - (n_1/n)) / (1+n_1/n)^2$, where $n_1$ = number responding and n = total number. We replaced $n_1/n$ with Column (2) since the quantities are the same.

**Column (4):** Equal to [1/Column (2)][1 + Column (3)]. The design effect (DEFF) is a ratio by which variance should be increased when using the imputed data. The factor 1/[Column (2)] is included to correct for the fact that standard variance calculations software treats the imputed values as real, thus overestimating the sample size. See Report No. 4 for more information on the use of "DEFF" in a more general context.

**Column (5):** The square root of DEFF.

## Table 4.3 — Multiple Imputation of Earnings by Size (Percentage Distribution)

**TOTAL**

| Earnings | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percentage Distribution (% Based on Weighted Frequencies) | | | | | | | | | | | | | |
| Overall | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.000 | 0.000 | 0.000 | 0.000 |
| No Earnings | 0.20 | 0.20 | 0.21 | 0.14 | 0.21 | 0.19 | 0.24 | 0.18 | 0.22 | 0.19 | 0.029 | 0.033 | 1.789 | 1.338 |
| $0–5K | 11.80 | 11.78 | 11.78 | 11.67 | 11.85 | 11.80 | 11.85 | 11.81 | 11.87 | 11.76 | 0.063 | 0.259 | 1.059 | 1.029 |
| $5K–10K | 10.32 | 10.35 | 10.32 | 10.40 | 10.43 | 10.31 | 10.20 | 10.31 | 10.27 | 10.32 | 0.071 | 0.259 | 1.075 | 1.037 |
| $10K–15K | 11.62 | 11.54 | 11.64 | 11.83 | 11.66 | 11.59 | 11.56 | 11.62 | 11.61 | 11.57 | 0.089 | 0.273 | 1.106 | 1.052 |
| $15K–20K | 11.01 | 11.06 | 10.96 | 11.27 | 10.91 | 10.92 | 11.15 | 10.91 | 11.06 | 10.88 | 0.139 | 0.274 | 1.259 | 1.122 |
| $20K–30K | 18.66 | 18.63 | 18.61 | 18.52 | 18.65 | 18.74 | 18.58 | 18.59 | 18.74 | 18.90 | 0.122 | 0.360 | 1.116 | 1.056 |
| $30K–50K | 23.23 | 23.32 | 23.19 | 23.23 | 23.28 | 23.26 | 23.16 | 23.29 | 23.04 | 23.29 | 0.093 | 0.402 | 1.054 | 1.026 |
| More than $50K | 13.15 | 13.12 | 13.29 | 12.95 | 13.02 | 13.19 | 13.25 | 13.29 | 13.19 | 13.07 | 0.128 | 0.246 | 1.269 | 1.127 |

**PARTNER**

| Earnings | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percentage Distribution (% Based on Weighted Frequencies) | | | | | | | | | | | | | |
| Overall | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.000 | 0.000 | 0.000 | 0.000 |
| No Earnings | 0.25 | 0.26 | 0.23 | 0.23 | 0.29 | 0.24 | 0.27 | 0.21 | 0.28 | 0.26 | 0.029 | 0.064 | 1.203 | 1.097 |
| $0–5K | 5.58 | 5.56 | 5.59 | 5.63 | 5.53 | 5.59 | 5.60 | 5.62 | 5.53 | 5.55 | 0.038 | 0.342 | 1.013 | 1.006 |
| $5K–10K | 6.69 | 6.70 | 6.75 | 6.62 | 6.81 | 6.61 | 6.58 | 6.74 | 6.67 | 6.75 | 0.083 | 0.341 | 1.059 | 1.029 |
| $10K–15K | 9.46 | 9.48 | 9.22 | 9.62 | 9.57 | 9.55 | 9.24 | 9.51 | 9.46 | 9.45 | 0.146 | 0.391 | 1.140 | 1.068 |
| $15K–20K | 10.04 | 10.04 | 10.13 | 10.32 | 10.03 | 9.82 | 10.27 | 10.04 | 9.97 | 9.76 | 0.194 | 0.390 | 1.249 | 1.117 |
| $20K–30K | 20.19 | 20.12 | 20.14 | 20.04 | 20.38 | 20.20 | 20.08 | 19.94 | 20.46 | 20.32 | 0.177 | 0.593 | 1.089 | 1.044 |
| $30K–50K | 28.56 | 28.60 | 28.42 | 28.68 | 28.56 | 28.37 | 28.56 | 28.77 | 28.45 | 28.63 | 0.135 | 0.654 | 1.043 | 1.021 |
| More than $50K | 19.26 | 19.23 | 19.52 | 18.86 | 19.05 | 19.61 | 19.40 | 19.17 | 19.18 | 19.30 | 0.244 | 0.529 | 1.213 | 1.102 |

**OPTION B**

| Earnings | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percentage Distribution (% Based on Weighted Frequencies) | | | | | | | | | | | | | |
| Overall | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.000 | 0.000 | 0.000 | 0.000 |
| No Earnings | 0.14 | 0.15 | 0.14 | 0.06 | 0.14 | 0.16 | 0.19 | 0.12 | 0.16 | 0.12 | 0.039 | 0.053 | 1.542 | 1.242 |
| $0–5K | 15.44 | 15.33 | 15.37 | 15.20 | 15.59 | 15.42 | 15.51 | 15.45 | 15.64 | 15.41 | 0.142 | 0.447 | 1.101 | 1.049 |
| $5K–10K | 12.31 | 12.44 | 12.25 | 12.45 | 12.53 | 12.35 | 12.18 | 12.20 | 12.16 | 12.25 | 0.142 | 0.511 | 1.077 | 1.038 |
| $10K–15K | 12.28 | 12.09 | 12.41 | 12.54 | 12.27 | 12.11 | 12.30 | 12.29 | 12.30 | 12.17 | 0.153 | 0.497 | 1.094 | 1.046 |
| $15K–20K | 11.24 | 11.35 | 11.16 | 11.54 | 11.04 | 11.18 | 11.39 | 11.03 | 11.27 | 11.22 | 0.173 | 0.496 | 1.122 | 1.059 |
| $20K–30K | 17.57 | 17.57 | 17.49 | 17.33 | 17.48 | 17.64 | 17.49 | 17.55 | 17.67 | 17.88 | 0.162 | 0.589 | 1.076 | 1.037 |
| $30K–50K | 20.65 | 20.77 | 20.71 | 20.61 | 20.70 | 20.92 | 20.40 | 20.68 | 20.35 | 20.76 | 0.190 | 0.565 | 1.113 | 1.055 |
| More than $50K | 10.37 | 10.29 | 10.45 | 10.27 | 10.24 | 10.21 | 10.53 | 10.67 | 10.44 | 10.20 | 0.172 | 0.406 | 1.180 | 1.086 |

**MKA's**

| Earnings | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Percentage Distribution (% Based on Weighted Frequencies) | | | | | | | | | | | | | |
| Overall | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 0.000 | 0.000 | 0.000 | 0.000 |
| No Earnings | 0.24 | 0.22 | 0.33 | 0.17 | 0.21 | 0.19 | 0.28 | 0.25 | 0.23 | 0.25 | 0.051 | 0.062 | 1.683 | 1.297 |
| $0–5K | 13.92 | 14.10 | 13.95 | 13.76 | 13.92 | 13.92 | 13.99 | 13.87 | 13.92 | 13.85 | 0.100 | 0.431 | 1.053 | 1.026 |
| $5K–10K | 11.84 | 11.69 | 11.84 | 12.03 | 11.69 | 11.82 | 11.71 | 11.94 | 11.93 | 11.87 | 0.128 | 0.421 | 1.092 | 1.045 |
| $10K–15K | 13.64 | 13.60 | 13.80 | 13.76 | 13.66 | 13.68 | 13.64 | 13.50 | 13.50 | 13.64 | 0.108 | 0.441 | 1.060 | 1.030 |
| $15K–20K | 12.05 | 12.04 | 11.82 | 12.18 | 12.03 | 12.07 | 12.03 | 12.01 | 12.33 | 11.93 | 0.152 | 0.421 | 1.131 | 1.063 |
| $20K–30K | 18.57 | 18.48 | 18.55 | 18.61 | 18.44 | 18.76 | 18.51 | 18.66 | 18.31 | 18.83 | 0.172 | 0.554 | 1.097 | 1.047 |
| $30K–50K | 20.29 | 20.39 | 20.21 | 20.19 | 20.55 | 20.16 | 20.50 | 20.18 | 20.19 | 20.27 | 0.157 | 0.595 | 1.069 | 1.034 |
| More than $50K | 9.45 | 9.48 | 9.51 | 9.30 | 9.49 | 9.40 | 9.33 | 9.59 | 9.60 | 9.36 | 0.114 | 0.469 | 1.059 | 1.029 |

Note: See text for definition of terms.

## Table 4.4 — Multiple Imputation of Earnings by Size

**TOTAL**

| Earnings | Average Earnings (In Dollars) | | | | | | | | | | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | | | | |
| Overall | 27800 | 27848 | 28034 | 27574 | 27691 | 27842 | 27899 | 27881 | 27658 | 27778 | 148.6 | 256.32 | 1.336 | 1.156 |
| No Earnings | -5554 | -5493 | -5158 | -6095 | -5842 | -5601 | -6174 | -4944 | -5379 | -5296 | 439.7 | 681.37 | 1.416 | 1.190 |
| $0–5K | 2107 | 2112 | 2105 | 2112 | 2108 | 2104 | 2112 | 2104 | 2109 | 2100 | 4.6 | 34.10 | 1.018 | 1.009 |
| $5K–10K | 7151 | 7168 | 7147 | 7142 | 7150 | 7144 | 7149 | 7140 | 7163 | 7152 | 9.9 | 40.19 | 1.061 | 1.030 |
| $10K–15K | 11938 | 11933 | 11929 | 11958 | 11951 | 11921 | 11939 | 11950 | 11914 | 11947 | 15.7 | 37.59 | 1.174 | 1.083 |
| $15K–20K | 16872 | 16875 | 16855 | 16907 | 16868 | 16877 | 16877 | 16860 | 16865 | 16866 | 15.9 | 43.73 | 1.132 | 1.064 |
| $20K–30K | 24047 | 24042 | 24075 | 24034 | 24048 | 24038 | 24054 | 24014 | 24064 | 24051 | 18.6 | 55.53 | 1.113 | 1.055 |
| $30K–50K | 37299 | 37323 | 37268 | 37304 | 37311 | 37298 | 37303 | 37316 | 37273 | 37294 | 19.4 | 105.35 | 1.034 | 1.017 |
| More than $50K | 79217 | 79619 | 80028 | 78464 | 79104 | 79312 | 79717 | 79154 | 78379 | 79174 | 573.3 | 1500.84 | 1.146 | 1.070 |

**PARTNER**

| Earnings | Average Earnings (In Dollars) | | | | | | | | | | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | | | | |
| Overall | 34178 | 34279 | 34295 | 33882 | 34024 | 34301 | 34417 | 34070 | 34174 | 34158 | 174.6 | 468.65 | 1.139 | 1.067 |
| No Earnings | -6256 | -5654 | -6119 | -6921 | -6571 | -5835 | -6371 | -5854 | -6465 | -6513 | 438.5 | 776.54 | 1.319 | 1.148 |
| $0–5K | 2192 | 2204 | 2133 | 2220 | 2197 | 2166 | 2210 | 2213 | 2175 | 2207 | 29.8 | 79.97 | 1.139 | 1.067 |
| $5K–10K | 7260 | 7292 | 7285 | 7186 | 7282 | 7259 | 7281 | 7217 | 7287 | 7247 | 38.7 | 64.82 | 1.357 | 1.165 |
| $10K–15K | 11969 | 11938 | 11959 | 11984 | 11969 | 11942 | 11990 | 12009 | 11960 | 11972 | 24.0 | 59.15 | 1.165 | 1.079 |
| $15K–20K | 16859 | 16891 | 16809 | 16897 | 16848 | 16863 | 16892 | 16839 | 16852 | 16843 | 31.0 | 66.23 | 1.218 | 1.104 |
| $20K–30K | 24043 | 24066 | 24032 | 23993 | 24056 | 24059 | 24034 | 24039 | 24101 | 24007 | 34.0 | 82.33 | 1.171 | 1.082 |
| $30K–50K | 37218 | 37288 | 37208 | 37223 | 37204 | 37157 | 37221 | 37265 | 37193 | 37207 | 40.7 | 153.77 | 1.070 | 1.034 |
| More than $50K | 79322 | 79804 | 79310 | 79091 | 78976 | 79109 | 80074 | 78879 | 79582 | 79070 | 432.0 | 1711.78 | 1.064 | 1.031 |

**OPTION B**

| Earnings | Average Earnings (In Dollars) | | | | | | | | | | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | | | | |
| Overall | 24728 | 24741 | 24965 | 24466 | 24633 | 24756 | 24791 | 24963 | 24462 | 24772 | 192.7 | 396.62 | 1.236 | 1.112 |
| No Earnings | -6407 | -6700 | -6451 | -7299 | -6099 | -6825 | -7145 | -5639 | -5760 | -5747 | 660.6 | 1259.99 | 1.275 | 1.129 |
| $0–5K | 2136 | 2136 | 2150 | 2121 | 2144 | 2132 | 2140 | 2129 | 2150 | 2123 | 11.4 | 52.65 | 1.047 | 1.023 |
| $5K–10K | 7131 | 7149 | 7116 | 7128 | 7126 | 7133 | 7113 | 7125 | 7134 | 7151 | 13.7 | 67.54 | 1.041 | 1.020 |
| $10K–15K | 11932 | 11941 | 11920 | 11957 | 11957 | 11921 | 11918 | 11939 | 11886 | 11946 | 24.0 | 63.37 | 1.144 | 1.069 |
| $15K–20K | 16914 | 16897 | 16902 | 16961 | 16912 | 16920 | 16902 | 16894 | 16917 | 16921 | 21.4 | 59.13 | 1.130 | 1.063 |
| $20K–30K | 24062 | 24019 | 24133 | 24043 | 24058 | 24032 | 24089 | 24016 | 24057 | 24111 | 43.2 | 100.02 | 1.187 | 1.089 |
| $30K–50K | 37411 | 37368 | 37346 | 37392 | 37458 | 37524 | 37363 | 37432 | 37380 | 37439 | 60.0 | 204.76 | 1.086 | 1.042 |
| More than $50K | 79167 | 79578 | 80851 | 77205 | 79298 | 79612 | 79578 | 79504 | 77052 | 79827 | 1304.9 | 2641.05 | 1.244 | 1.115 |

**MKA's**

| Earnings | Average Earnings (In Dollars) | | | | | | | | | | Impute Error | Sampling Error | DEFF | DEFT |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Overall | Iteration 1 | Iteration 2 | Iteration 3 | Iteration 4 | Iteration 5 | Iteration 6 | Iteration 7 | Iteration 8 | Iteration 9 | | | | |
| Overall | 24244 | 24296 | 24372 | 24216 | 24330 | 24206 | 24212 | 24313 | 24159 | 24095 | 94.2 | 553.83 | 1.029 | 1.014 |
| No Earnings | -3338 | -3467 | -2917 | -3477 | -3922 | -2977 | -4497 | -3100 | -2734 | -2954 | 600.6 | 940.33 | 1.408 | 1.187 |
| $0–5K | 1988 | 2002 | 1983 | 2022 | 1969 | 2001 | 1986 | 1977 | 1973 | 1983 | 17.8 | 48.77 | 1.133 | 1.064 |
| $5K–10K | 7098 | 7100 | 7093 | 7133 | 7084 | 7069 | 7110 | 7105 | 7116 | 7072 | 22.0 | 45.08 | 1.238 | 1.113 |
| $10K–15K | 11916 | 11912 | 11914 | 11933 | 11919 | 11898 | 11922 | 11907 | 11917 | 11923 | 10.6 | 46.97 | 1.051 | 1.025 |
| $15K–20K | 16808 | 16812 | 16826 | 16815 | 16809 | 16813 | 16808 | 16822 | 16782 | 16788 | 15.3 | 57.52 | 1.071 | 1.035 |
| $20K–30K | 24023 | 24047 | 24035 | 24085 | 24013 | 24014 | 24018 | 23970 | 24016 | 24005 | 33.2 | 86.37 | 1.148 | 1.071 |
| $30K–50K | 37236 | 37303 | 37233 | 37297 | 37233 | 37120 | 37356 | 37181 | 37224 | 37177 | 77.0 | 178.74 | 1.186 | 1.089 |
| More than $50K | 79007 | 79128 | 80445 | 79369 | 79069 | 79293 | 78886 | 79202 | 77634 | 78037 | 848.0 | 3926.94 | 1.047 | 1.023 |

Note: See text for definition of terms.

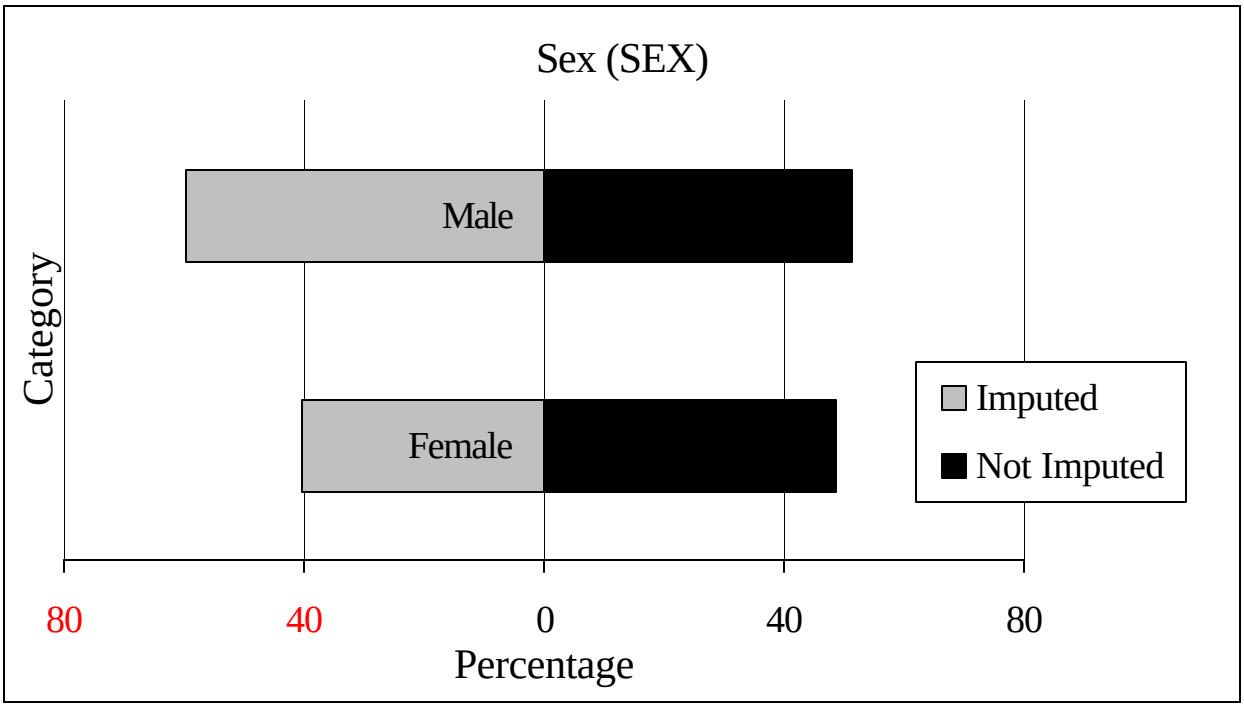
## Table 4.5 — Focal Child Public Use File — Age (AGE): Imputed, Not Imputed

| Age (years) | All Data | | | | Imputed Data | | | | Not Imputed Data | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Weighted | | Unweighted | | Weighted | | Unweighted | | Weighted | | Unweighted | |
| | N | % | n | % | N | % | n | % | N | % | n | % |
| 0 | 3669994 | 5.27 | 1705 | 5.06 | 0 | 0.00 | 0 | 0.00 | 3669994 | 5.29 | 1705 | 5.07 |
| 1 | 3281311 | 4.71 | 1668 | 4.95 | 0 | 0.00 | 0 | 0.00 | 3281311 | 4.73 | 1668 | 4.96 |
| 2 | 4529615 | 6.51 | 2062 | 6.12 | 177 | 0.06 | 1 | 1.82 | 4529438 | 6.53 | 2061 | 6.13 |
| 3 | 3587658 | 5.15 | 2194 | 6.51 | 0 | 0.00 | 0 | 0.00 | 3587658 | 5.18 | 2194 | 6.52 |
| 4 | 4122807 | 5.92 | 2316 | 6.87 | 2239 | 0.75 | 2 | 3.64 | 4120568 | 5.94 | 2314 | 6.88 |
| 5 | 4412736 | 6.34 | 2375 | 7.05 | 45132 | 15.11 | 9 | 16.36 | 4367605 | 6.30 | 2366 | 7.03 |
| 6 | 3990203 | 5.73 | 2059 | 6.11 | 3473 | 1.16 | 3 | 5.45 | 3986731 | 5.75 | 2056 | 6.11 |
| 7 | 3862404 | 5.55 | 1995 | 5.92 | 20737 | 6.94 | 3 | 5.45 | 3841667 | 5.54 | 1992 | 5.92 |
| 8 | 4244123 | 6.10 | 1911 | 5.67 | 2885 | 0.97 | 3 | 5.45 | 4241238 | 6.12 | 1908 | 5.67 |
| 9 | 3721015 | 5.34 | 1690 | 5.01 | 10478 | 3.51 | 3 | 5.45 | 3710536 | 5.35 | 1687 | 5.01 |
| 10 | 3959722 | 5.69 | 1713 | 5.08 | 55694 | 18.64 | 6 | 10.91 | 3904028 | 5.63 | 1707 | 5.07 |
| 11 | 3731166 | 5.36 | 1581 | 4.69 | 49025 | 16.41 | 4 | 7.27 | 3682141 | 5.31 | 1577 | 4.69 |
| 12 | 3796779 | 5.45 | 1668 | 4.95 | 4052 | 1.36 | 2 | 3.64 | 3792727 | 5.47 | 1666 | 4.95 |
| 13 | 3425021 | 4.92 | 1603 | 4.76 | 6640 | 2.22 | 3 | 5.45 | 3418381 | 4.93 | 1600 | 4.76 |
| 14 | 4176762 | 6.00 | 1735 | 5.15 | 53864 | 18.03 | 6 | 10.91 | 4122899 | 5.95 | 1729 | 5.14 |
| 15 | 3555769 | 5.11 | 1696 | 5.03 | 4365 | 1.46 | 7 | 12.73 | 3551403 | 5.12 | 1689 | 5.02 |
| 16 | 3868288 | 5.56 | 1923 | 5.71 | 38427 | 12.86 | 2 | 3.64 | 3829860 | 5.53 | 1921 | 5.71 |
| 17 | 3681701 | 5.29 | 1809 | 5.37 | 1581 | 0.53 | 1 | 1.82 | 3680120 | 5.31 | 1808 | 5.37 |

Note: See text for definition of terms.

**Table 4.6 — Focal Child Public Use File — Gender (SEX): Imputed, Not Imputed**

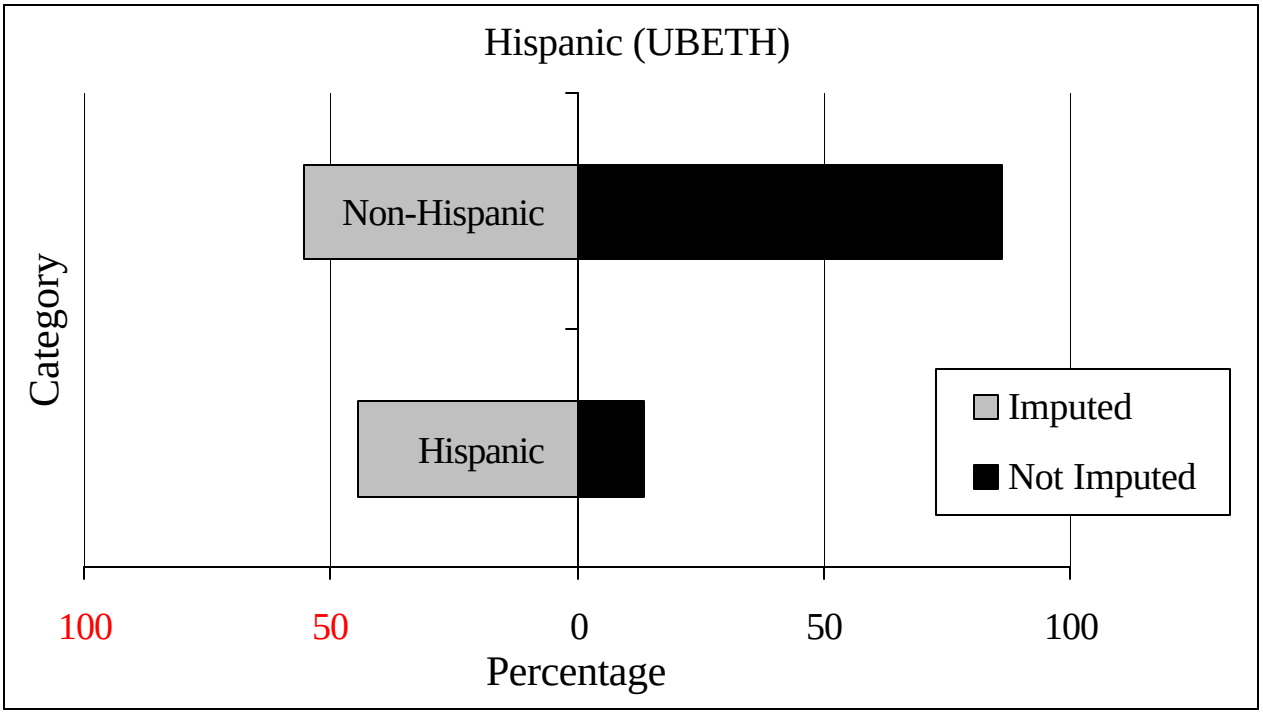
| All Data | | | | |
|---|---|---|---|---|
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 33982277 | 48.81 | 16343 | 48.49 |
| Male | 35634797 | 51.19 | 17360 | 51.51 |
| **Imputed Data** | | | | |
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 55773 | 40.28 | 19 | 51.35 |
| Male | 82698 | 59.72 | 18 | 48.65 |
| **Not Imputed Data** | | | | |
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 33926503 | 48.83 | 16324 | 48.49 |
| Male | 35552099 | 51.17 | 17342 | 51.51 |




Note: This graph illustrates the ratio of females and males for imputed data on the left side. On the right side the graph illustrates the male to female ratio for not imputed data. All data is weighted.

**Table 4.7 — Focal Child Public Use File — Hispanic (UBETH): Imputed, Not Imputed**

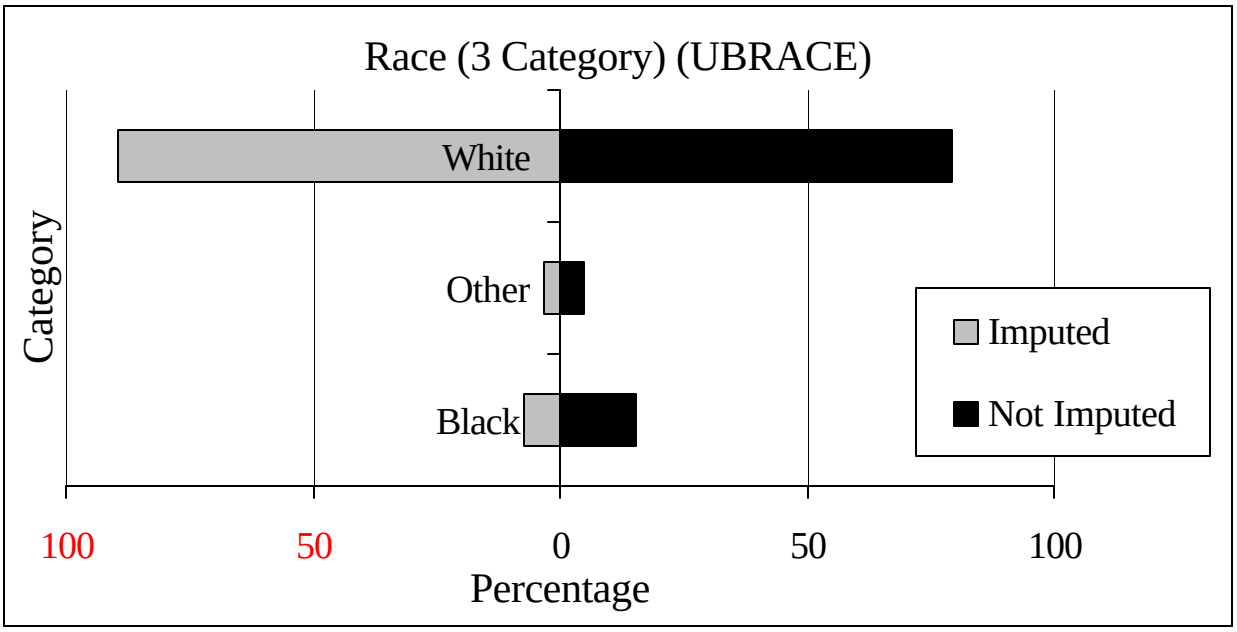
| All Data | | | | |
|---|---|---|---|---|
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 10030337 | 14.41 | 4984 | 14.79 |
| Non-Hispanic | 59586737 | 85.59 | 28719 | 85.21 |
| **Imputed Data** | | | | |
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 783028 | 44.44 | 555 | 53.16 |
| Non-Hispanic | 978861 | 55.56 | 489 | 46.84 |
| **Not Imputed Data** | | | | |
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 9247309 | 13.63 | 4429 | 13.56 |
| Non-Hispanic | 58607876 | 86.37 | 28230 | 86.44 |




Note: This graph illustrates the ratio of Hispanics to non-Hispanics for imputed data on the left side. The right side of the graph illustrates the Hispanic to non-Hispanic ratio for not imputed data. All data is weighted.

**Table 4.8 — Focal Child Public Use File — Race (3 category) (UBRACE): Imputed, Not Imputed**

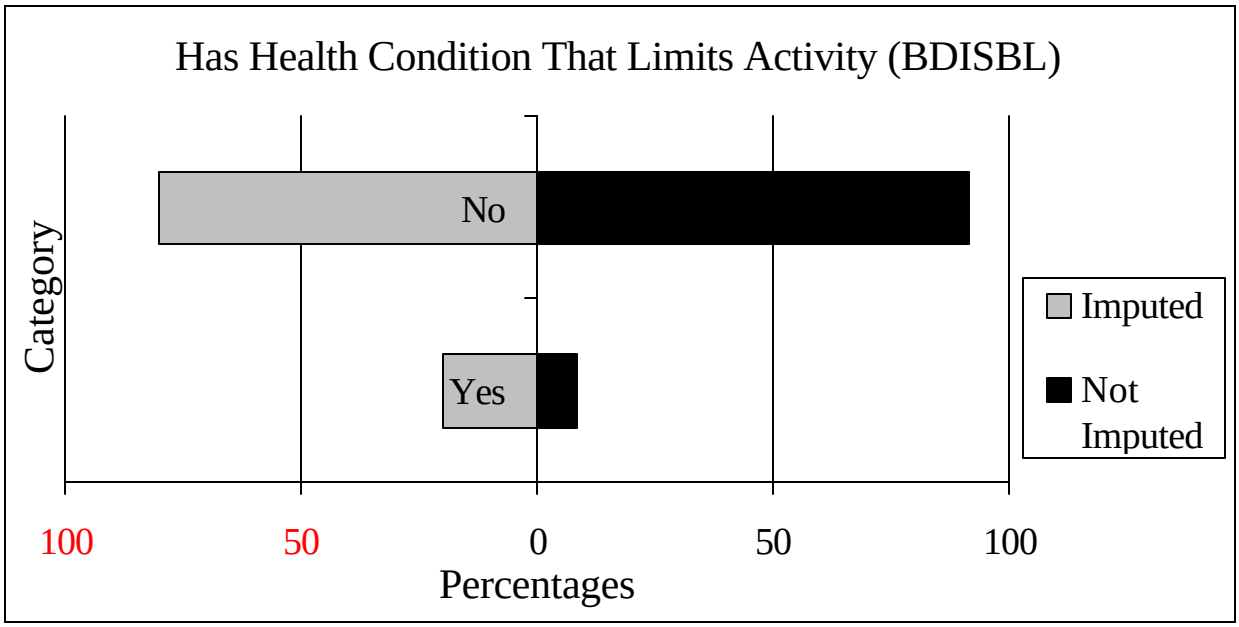
| All Data | | | | |
|---|---|---|---|---|
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 10659374 | 15.31 | 5269 | 15.63 |
| Other | 3487773 | 5.01 | 1423 | 4.22 |
| White | 55469927 | 79.68 | 27011 | 80.14 |
| **Imputed Data** | | | | |
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 128889 | 7.32 | 81 | 7.76 |
| Other | 58968 | 3.35 | 29 | 2.78 |
| White | 1574032 | 89.34 | 934 | 89.46 |
| **Not Imputed Data** | | | | |
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 10530484 | 15.52 | 5188 | 15.89 |
| Other | 3428805 | 5.05 | 1394 | 4.27 |
| White | 53895895 | 79.43 | 26077 | 79.85 |




Note: This graph illustrates the ratio of Hispanics to non-Hispanics for imputed data on the left side. The right side illustrates that ratio for not imputed data. All data is weighted.

**Table 4.9 — Focal Child Public Use File — Has Health Condition that Limits Activity (BDISBL): Imputed, Not Imputed**

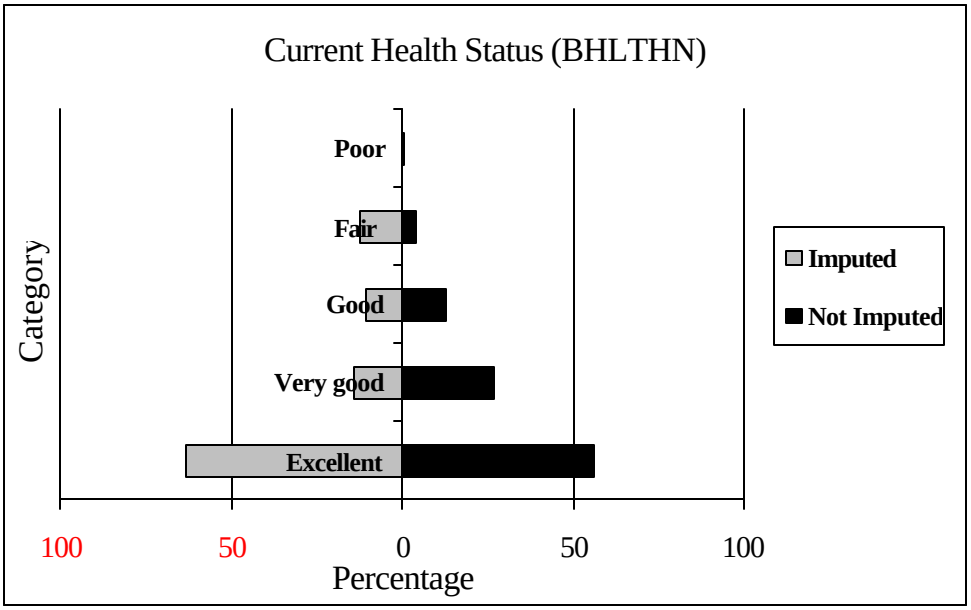
| All Data | | | | |
|---|---|---|---|---|
| BDISBL | Weighted | | Unweighted | |
| | N | % | n | % |
| Yes | 5786365 | 8.31 | 3130 | 9.29 |
| No | 63830709 | 91.69 | 30573 | 90.71 |
| **Imputed Data** | | | | |
| BDISBL | Weighted | | Unweighted | |
| | N | % | n | % |
| Yes | 55654 | 19.91 | 25 | 17.36 |
| No | 223840 | 80.09 | 119 | 82.64 |
| **Not Imputed Data** | | | | |
| BDISBL | Weighted | | Unweighted | |
| | N | % | n | % |
| Yes | 5730711 | 8.26 | 3105 | 9.25 |
| No | 63606869 | 91.74 | 30454 | 90.75 |




Note: This graph illustrates the ratio of people who answered yes to those who answered no for imputed data on the left side. The right side illustrates the same ratio for not imputed data. All data is weighted.

## Table 4.10 — Focal Child Public Use File — Current Health Status (BHLTHN): Imputed, Not Imputed

| All Data | | | | |
|---|---|---|---|---|
| BHLTHN | Weighted | | Unweighted | |
| | N | % | N | % |
| Excellent | 39020510 | 56.05 | 18165 | 53.90 |
| Very good | 18534000 | 26.62 | 9363 | 27.78 |
| Good | 8888084 | 12.77 | 4561 | 13.53 |
| Fair | 2814702 | 4.04 | 1414 | 4.20 |
| Poor | 359777 | 0.52 | 200 | 0.59 |
| **Imputed Data** | | | | |
| BHLTHN | Weighted | | Unweighted | |
| | N | % | N | % |
| Excellent | 30969 | 62.95 | 18 | 58.06 |
| Very good | 7029 | 14.29 | 6 | 19.35 |
| Good | 5111 | 10.39 | 4 | 12.90 |
| Fair | 6086 | 12.37 | 3 | 9.68 |
| Poor | 0 | 0.00 | 0 | 0.00 |
| **Not Imputed Data** | | | | |
| BHLTHN | Weighted | | Unweighted | |
| | N | % | N | % |
| Excellent | 38989542 | 56.05 | 18147 | 53.89 |
| Very good | 18526970 | 26.63 | 9357 | 27.79 |
| Good | 8882973 | 12.77 | 4557 | 13.53 |
| Fair | 2808616 | 4.04 | 1411 | 4.19 |
| Poor | 359777 | 0.52 | 200 | 0.59 |




Note: This graph illustrates the proportion of people who answered poor, fair, good, very good, and excellent for imputed data on the left side. On the right side, the graph illustrates the proportion of people who answered poor, fair, good, very good, and excellent for not imputed data. All data is weighted.

**Table 4.11 — Focal Child Public Use File — Current Health Compared to 12 Months Ago (BHLTHP): Imputed, Not Imputed**

**All Data**

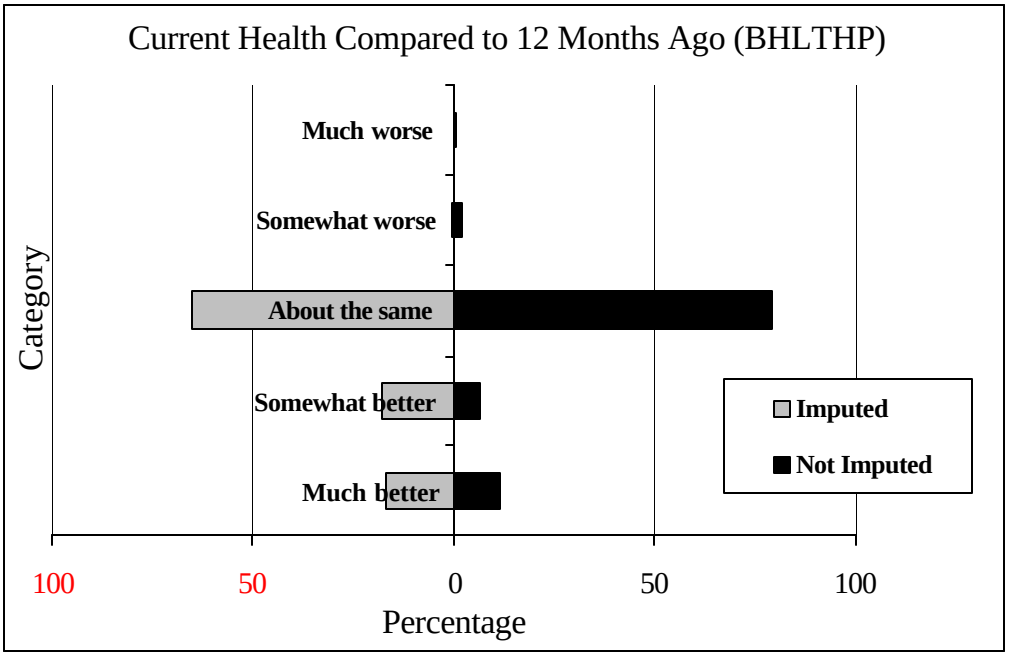
| BHLTHP | Weighted | | Unweighted | |
|---|---|---|---|---|
| | N | % | n | % |
| Inapplicable | 6951305 | 0.00 | 3373 | 0.00 |
| Much better | 7280967 | 11.62 | 3759 | 12.39 |
| Somewhat better | 4292131 | 6.85 | 2309 | 7.61 |
| About the same | 49687500 | 79.29 | 23558 | 77.67 |
| Somewhat worse | 1263310 | 2.02 | 620 | 2.04 |
| Much worse | 141862 | 0.23 | 84 | 0.28 |

**Imputed Data**

| BHLTHP | Weighted | | Unweighted | |
|---|---|---|---|---|
| | N | % | n | % |
| Inapplicable | 0 | 0.00 | 0 | 0.00 |
| Much better | 21906 | 16.66 | 8 | 12.50 |
| Somewhat better | 23233 | 17.67 | 8 | 12.50 |
| About the same | 85747 | 65.22 | 47 | 73.44 |
| Somewhat worse | 597 | 0.45 | 1 | 1.56 |
| Much worse | 0 | 0.00 | 0 | 0.00 |

**Not Imputed Data**

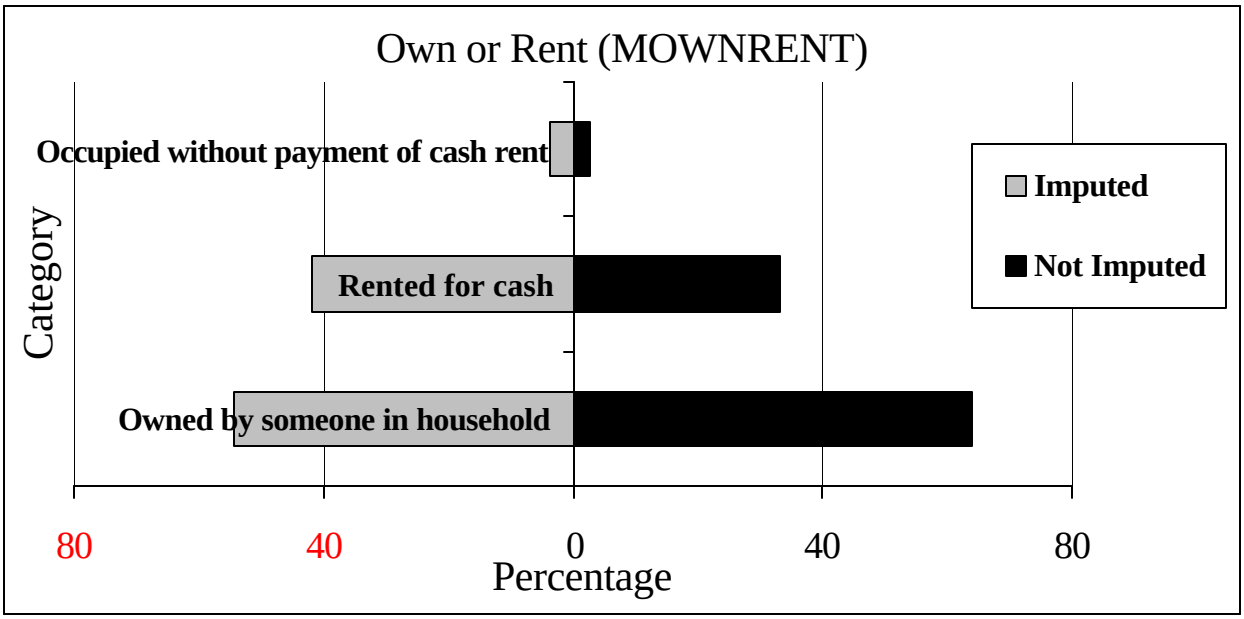
| BHLTHP | Weighted | | Unweighted | |
|---|---|---|---|---|
| | N | % | n | % |
| Inapplicable | 6951305 | 0.00 | 3373 | 0.00 |
| Much better | 7259061 | 11.61 | 3751 | 12.39 |
| Somewhat better | 4268898 | 6.83 | 2301 | 7.60 |
| About the same | 49601753 | 79.32 | 23511 | 77.68 |
| Somewhat worse | 1262712 | 2.02 | 619 | 2.05 |
| Much worse | 141862 | 0.23 | 84 | 0.28 |




Note: This graph illustrates the proportion of people who answered much worse, somewhat worse, about the same, somewhat better, and much better for imputed data on the left side. On the right side, the graph illustrates the proportion of people who answered much worse, somewhat worse, about the same, somewhat better, and much better for not imputed data. All data is weighted.

**Table 4.12 — Focal Child Public Use File — Own or Rent (MOWNRENT): Imputed, Not Imputed**

| All Data | | | | |
|---|---|---|---|---|
| MOWNRENT | Weighted | | Unweighted | |
| | N | % | n | % |
| Owned by someone in household | 44547625 | 63.99 | 21161 | 62.79 |
| Rented for cash | 23176894 | 33.29 | 11629 | 34.50 |
| Occupied without payment of cash rent | 1892554 | 2.72 | 913 | 2.71 |
| **Imputed Data** | | | | |
| MOWNRENT | Weighted | | Unweighted | |
| | N | % | n | % |
| Owned by someone in household | 389338 | 54.42 | 222 | 60.82 |
| Rented for cash | 299456 | 41.86 | 131 | 35.89 |
| Occupied without payment of cash rent | 26605 | 3.72 | 12 | 3.29 |
| **Not Imputed Data** | | | | |
| MOWNRENT | Weighted | | Unweighted | |
| | N | % | n | % |
| Owned by someone in household | 44158287 | 64.09 | 20939 | 62.81 |
| Rented for cash | 22877438 | 33.20 | 11498 | 34.49 |
| Occupied without payment of cash rent | 1865949 | 2.71 | 901 | 2.70 |



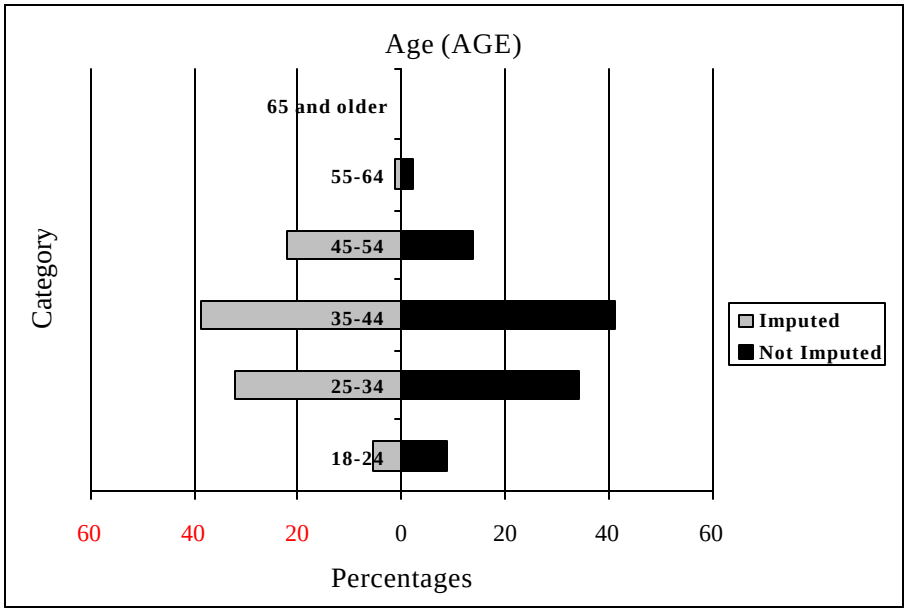

Note: This graph illustrates the ratio of people who occupy without payment of cash rent, rent for cash, and live in a place owned by someone in their household for imputed data on the left side and for not imputed data on the right side. All data is weighted.

## Table 4.13 — MKA Public Use File — Age (AGE): Imputed, Not Imputed

| All Data | | | | |
|---|---|---|---|---|
| AGE | Weighted | | Unweighted | |
| | N | % | n | % |
| 16 | 0 | 0.00 | 11 | 0.04 |
| 17 | 0 | 0.00 | 34 | 0.12 |
| 18–24 | 3178338 | 8.52 | 2094 | 7.64 |
| 25–34 | 12702872 | 34.05 | 8859 | 32.33 |
| 35–44 | 15425902 | 41.35 | 11408 | 41.64 |
| 45–54 | 5204374 | 13.95 | 4185 | 15.27 |
| 55–64 | 789727 | 2.12 | 616 | 2.25 |
| 65 and older | 0 | 0.00 | 191 | 0.70 |

| Imputed Data | | | | |
|---|---|---|---|---|
| AGE | Weighted | | Unweighted | |
| | N | % | n | % |
| 16 | 0 | 0.00 | 0 | 0.00 |
| 17 | 0 | 0.00 | 0 | 0.00 |
| 18–24 | 17172 | 5.45 | 11 | 5.82 |
| 25–34 | 101543 | 32.23 | 48 | 25.40 |
| 35–44 | 121898 | 38.69 | 82 | 43.39 |
| 45–54 | 70115 | 22.25 | 43 | 22.75 |
| 55–64 | 4327 | 1.37 | 5 | 2.65 |
| 65 and older | 0 | 0.00 | 0 | 0.00 |

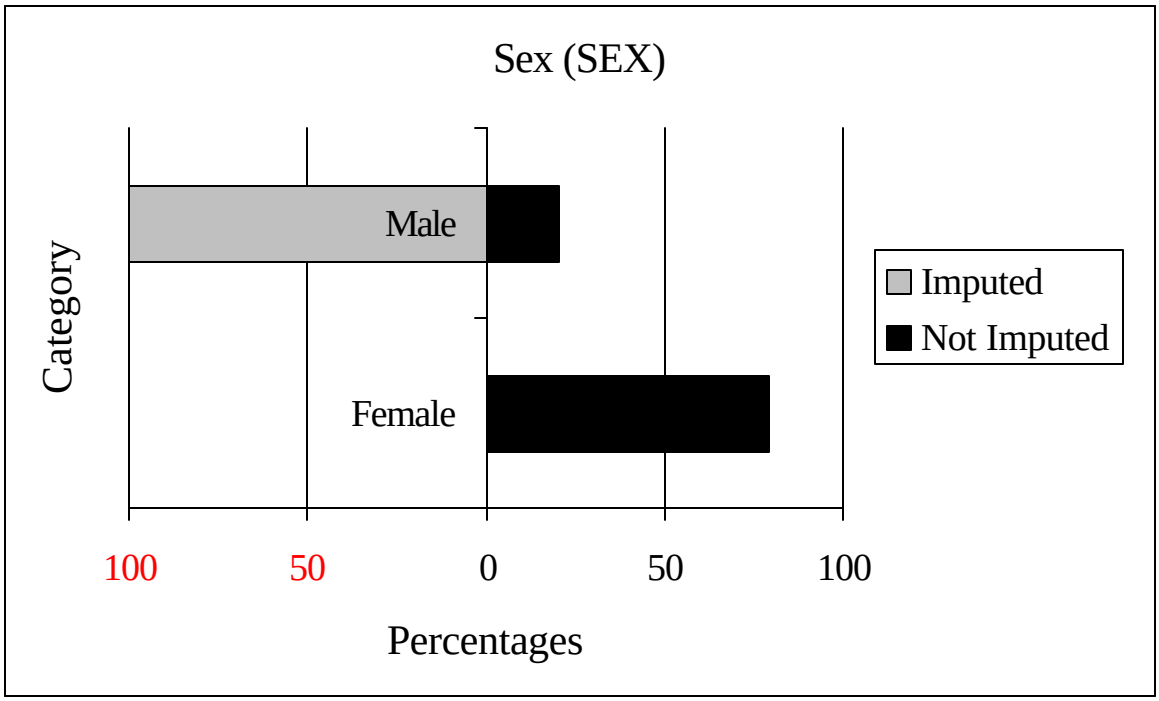
| Not Imputed Data | | | | |
|---|---|---|---|---|
| AGE | Weighted | | Unweighted | |
| | N | % | n | % |
| 16 | 0 | 0.00 | 11 | 0.04 |
| 17 | 0 | 0.00 | 34 | 0.12 |
| 18–24 | 3161166 | 8.55 | 2083 | 7.66 |
| 25–34 | 12601330 | 34.07 | 8811 | 32.38 |
| 35–44 | 15304004 | 41.38 | 11326 | 41.63 |
| 45–54 | 5134259 | 13.88 | 4142 | 15.22 |
| 55–64 | 785400 | 2.12 | 611 | 2.25 |
| 65 and older | 0 | 0.00 | 191 | 0.70 |




Note: This table illustrates the proportion of different age brackets for imputed data on the left and not imputed data on the right. All data is weighted.

**Table 4.14 — MKA Public Use File — Sex (SEX): Imputed, Not Imputed**

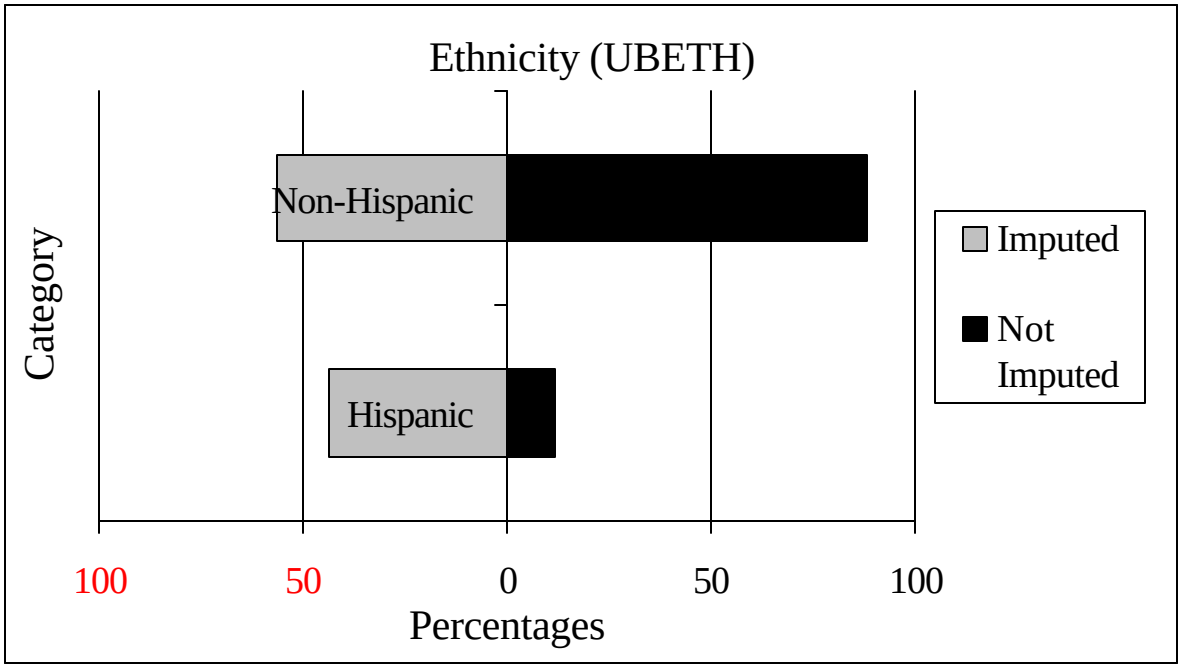
| All Data | | | | |
|---|---|---|---|---|
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 29632671 | 79.44 | 22275 | 81.30 |
| Male | 7668542 | 20.56 | 5123 | 18.70 |
| **Imputed Data** | | | | |
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 0 | 0.00 | 0 | 0.00 |
| Male | 814 | 100.00 | 2 | 100.00 |
| **Not Imputed Data** | | | | |
| SEX | Weighted | | Unweighted | |
| | N | % | n | % |
| Female | 29632671 | 79.44 | 22275 | 81.31 |
| Male | 7667728 | 20.56 | 5121 | 18.69 |




Note: This graph illustrates the ratio of males and females for imputed data on the left side. On the right side the graph illustrates the male to female ratio for not imputed data. All data is weighted.

**Table 4.15 — MKA Public Use File — Ethnicity (UBETH): Imputed, Not Imputed**

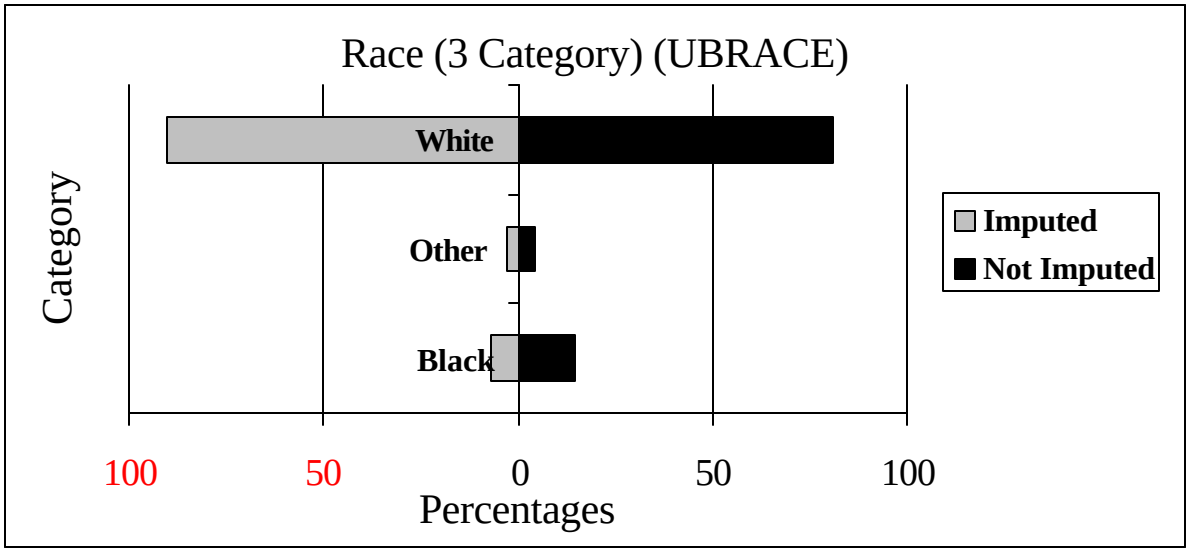
| All Data | | | | |
|---|---|---|---|---|
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 4762963 | 12.77 | 3519 | 12.84 |
| Non-Hispanic | 32538250 | 87.23 | 23879 | 87.16 |
| **Imputed Data** | | | | |
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 427239 | 43.42 | 440 | 53.46 |
| Non-Hispanic | 556661 | 56.58 | 383 | 46.54 |
| **Not Imputed Data** | | | | |
| UBETH | Weighted | | Unweighted | |
| | N | % | N | % |
| Hispanic | 4335724 | 11.94 | 3079 | 11.59 |
| Non-Hispanic | 31981588 | 88.06 | 23496 | 88.41 |




Note: This graph illustrates the ratio of Hispanics to non-Hispanics for imputed data on the left side. On the right side, the graph illustrates the Hispanic to non-Hispanic ratio for not imputed data. All data is weighted.

**Table 4.16 — MKA Public Use File — Race (3 Category) (UBRACE): Imputed, Not Imputed**

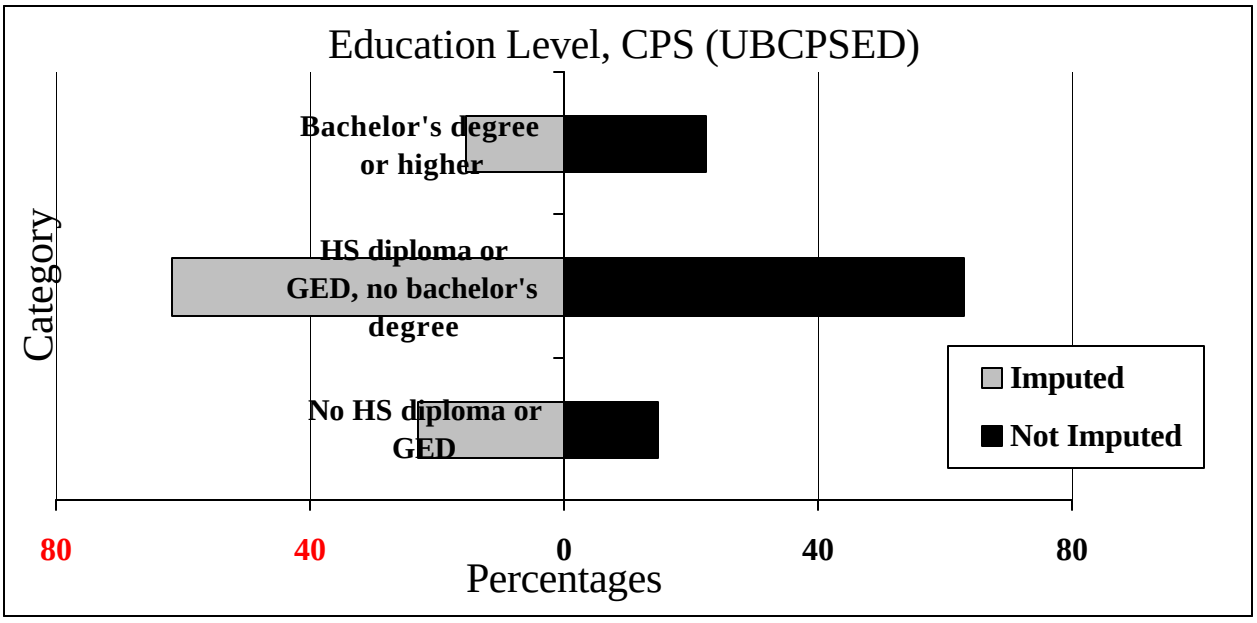
| All Data | | | | |
|---|---|---|---|---|
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 5418375 | 14.53 | 3959 | 14.45 |
| Other | 1633853 | 4.38 | 1094 | 3.99 |
| White | 30248985 | 81.09 | 22345 | 81.56 |
| **Imputed Data** | | | | |
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 68737 | 6.99 | 61 | 7.41 |
| Other | 27035 | 2.75 | 21 | 2.55 |
| White | 888128 | 90.27 | 741 | 90.04 |
| **Not Imputed Data** | | | | |
| UBRACE | Weighted | | Unweighted | |
| | N | % | N | % |
| Black | 5349637 | 14.73 | 3898 | 14.67 |
| Other | 1606818 | 4.42 | 1073 | 4.04 |
| White | 29360857 | 80.85 | 21604 | 81.29 |




Note: This graph illustrates the proportion of white, black, and other for imputed data on the left side. On the right side, the graph illustrates the proportion of white, black, and other for not imputed data. All data is weighted.

**Table 4.17 — MKA Public Use File — Education Level, CPS (UBCPSED): Imputed, Not Imputed**

| All Data | | | | |
|---|---|---|---|---|
| UBCPSED | Weighted | | Unweighted | |
| | N | % | N | % |
| No HS diploma or GED | 5521709 | 14.80 | 3692 | 13.48 |
| HS diploma or GED, no bachelor's degree | 23446592 | 62.86 | 17662 | 64.46 |
| Bachelor's degree or higher | 8332911 | 22.34 | 6044 | 22.06 |
| **Imputed Data** | | | | |
| UBCPSED | Weighted | | Unweighted | |
| | N | % | N | % |
| No HS diploma or GED | 47702 | 23.04 | 35 | 21.08 |
| HS diploma or GED, no bachelor's degree | 127562 | 61.62 | 112 | 67.47 |
| Bachelor's degree or higher | 31761 | 15.34 | 19 | 11.45 |
| **Not Imputed Data** | | | | |
| UBCPSED | Weighted | | Unweighted | |
| | N | % | N | % |
| No HS diploma or GED | 5474007 | 14.76 | 3657 | 13.43 |
| HS diploma or GED, no bachelor's degree | 23319030 | 62.86 | 17550 | 64.45 |
| Bachelor's degree or higher | 8301149 | 22.38 | 6025 | 22.12 |




Note: This graph illustrates education level for imputed data on the left side. On the right side, the graph illustrates education level for not imputed data. All data is weighted.

# References


Andrews, L., Long, S., and Marquis, S. (1997), "Data Cleaning Procedures for the 1993 Robert Wood Johnson Foundation Family Health Insurance Survey." Santa Monica, CA: RAND.

Bassi, F., and Luigi, F. (1997), "Estimators of Nonsampling Errors in Interview-Reinterview Supervised Surveys with Interpenetrated Assignments." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Batcher, M., and Scheuren, F. (1997), "CATI Site Management in a Survey of Service Quality." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Bethlehem, J. (1997), "Integrated Control Systems for Survey Processing." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Biemer, P., and Trewin, D. (1997), "A Review of Measurement Error Effects on the Analysis of Survey Data." *Survey Methods and Process Quality.* Eds. Lars
Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Carlson, B., Cox, B., and Bandeh, L. (1999), *"SAS Macros Useful in Imputing Missing Survey Data."* Princeton, NJ: Mathematica Policy Research, Inc.

Chapman, D. (1974), *An Investigation of Nonresponse Imputation Procedures for the Health and Nutrition Examination Survey*. Rockville, MD: Westat.

Coder, J. (1999), "Imputation of a Missing Survey Response." (Unpublished)

Colledge, M., and March, M. (1997), "Quality Policies, Standards, Guidelines, and Recommended Practices at National Statistical Agencies." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Cox, B., and Folsom, R. (1981), "An Evaluation of Weighted Hot Deck Imputation for Unreported Health Care Visits." Proceedings of the Section of Survey Research Methods, American Statistical Association, 1981: 412–417.

Dippo. (1997), "Survey Measurement and Process Improvement: Concepts and Integration." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Fellegi, I., and Holt, D. (1976), A Systematic Approach to Automatic Edit and Imputation. *Journal of the American Statistical Association*, 71, 353: 17–35.

Ford, B. (1983), "An Overview of Hot-Deck Procedures." *Incomplete Data in Sample Surveys, Vol. 2*. Burlington, MA: Academic Press, Inc.

Granquist, L. (1995), "Improving the Traditional Editing Process." *Business Survey Methods*. Edited by Cox, et al. Wiley and Sons, Inc.

Granquist, L., and Kovar, J. (1997), "Editing of Survey Data: How Much is Enough?" *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Greenlees, J., Reece, W., and Zeischang, K. (1982), "Imputation of Missing Values When the Probability of Response Depends Upon the Variable Being Imputed." *Journal of the American Statistical Association* 77: 251–261.

Groves, R., and Couper, M. (1998*), Nonresponse in Household Interview Surveys*. New York, NY: Wiley and Sons, Inc.

Hansen, M., and Hurwitz, W. (1946), "The Problem for Nonresponse in Sample Surveys." *Journal of the American Statistical Association*, 41.: 517–529.

Hansen, M., Hurwitz, W., and Madow, W. (1953*), Sample Survey Methods and Theory (Vol. 2)*: 129–141. New York, NY: Wiley and Sons, Inc.

Holt, T., ed. (1997), *Statistical Quality.* London: National Statistics Information and Library Service.

Jabine, T. (1994), *Quality Profile for SASS*. United States Department of Education, National Center for Education Statistics. Washington DC.

Kalton, G. (1983), *Research Report Series: Compensating for Missing Survey Data.* Ann Arbor: Survey Research Center, Institute for Social Research, University of Michigan.

King, G., Honaker, J. Joseph, A., and Scheve, K. (1998). "Listwise Deletion is Evil: What to Do About Missing Data in Political Science." Paper presented at the Annual Meetings of the American Political Science Association, Boston, MA: September 3-6.

Leeuw, E., and Collins, M. (1997), "Data Collection Methods and Survey Quality: An Overview." *Survey Methods and Process Quality*. Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Lessler, J., and Kalsbeek, W. (1992), N*onsampling Error in Surveys*. New York, NY: Wiley and Sons, Inc.

Little, R. J. A. 1986. "Survey Nonresponse Adjustments." *International Statistical Review* 54: 139–17.

Little, R., and Rubin, D. (1987), *Statistical Analysis with Missing Data.* New York, NY: Wiley and Sons, Inc.

Lyberg, L., and Kasprzyk, D. (1997), "Some Aspects of Post-Survey Processing." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Madow, W., and Olkin, I. (1983), *Incomplete Data in Sample Surveys, Volume 2: Proceedings of the Symposium.* New York, NY: Academic Press.

Madow, W., Nisselson, H. and Olkin, I. (1983), *Incomplete Data in Sample Surveys, Volume 1: Report and Case Studies.* New York, NY: Academic Press.

Madow, W., Olkin, I., and Rubin, D. (1983), *Incomplete Data in Sample Surveys, Volume 3: Theory and Bibliographies*. New York, NY: Academic Press.

Morganstein, D., and Marker, D. (1997), "Continuous Quality Improvement in Statistical Agencies." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Oh, H., and Scheuren, F. (1983), "Weighting Adjustment for Unit Nonresponse." *Incomplete Data in Sample Surveys*, Vol. 2. Burlington, MA: Academic Press, Inc.

Oh, H., and Scheuren, F. with H. Nisselson. (1980a), "Differential Bias Impacts of Alternative Census Bureau Hot Deck Procedures for Imputing Missing CPS Income Data." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1980: 416–420.

Oh, H., and Scheuren, F. with G. Shapiro. (1980b), "Estimating the Variance Impact of Missing CPS Income Data." *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 1980: 408–415.

Organizing Committee for the Data Editing Workshop and Exposition, Federal Committee on Statistical Methodology. (1996), *Statistical Policy Working Paper 25: Data Editing Workshop and Exposition.* Washington, DC: Statistical Policy Office.

Pol, F., and Laneheine, P. (1997), "Separating Change and Measurement Error in a Panel Surveys with and Application to Labor Market Data." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Rao, J., and Sitter, R. (1997), "Variance Estimation Under Stratified Two-Phase Sampling with Applications to Measurement Bias." *Survey Methods and Process Quality.* Eds. Lars Lyberg, et al. New York, NY: Wiley and Sons, Inc.

Rubin, D. (1977), "Formalizing Subjective Notions about the Effect of Nonrespondents in Sample Surveys." *Journal of the American Statistical Association*, 72: 538–543.

Rubin, D. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York, NY: Wiley and Sons, Inc.

Schafer, J. (1997), *Analysis of Incomplete Multivariate Data.* New York, NY: Chapman & Hall.

Scherkenbach, W. (1991), *Deming's Road to Continual Improvement*. Knoxville, TN: SPC Press.

Scheuren, F. (1976), *Preliminary Notes on the Partially Missing Data Problem—Some (Very) Elementary Considerations*. Delivered at the April 1976 meeting of the Social Security Administration's Statistical Methodology Group.

Scheuren, F. (1978), "Discussion of the Paper Entitled, Income Data Collection and Processing for the March Income Supplement to the Current Population Survey, by John Coder." *Proceedings of the Data Processing Workshop: Survey of Income and Program Participation*: Washington, DC: U.S. Department of Health, Education, and Welfare.

Schieber, S. (1978), "A Comparison of Three Alternative Techniques for Allocating Unreported Social Security Income on the Survey of Low-Income Aged and Disabled. American Statistical Association." *American Statistical Association, Proceedings of the Section on Survey research Methods*: 212-18.

Shao, J., Chen Y., and Chen Y. (1998), "Balanced Repeated Replication for Stratifies Multistage Survey Data Under Imputation." *Journal of the American Statistical Association*, 93: 818-831.

U.S. Census Burea. (1998), Current Report Series p-60, No. 198.

Welniak, E., and Coder, F. (1980), A Measure of the Bias in the March CPS Earnings Imputation System and Results of a Sample Bias Adjustment Procedure. *American Statistical Association, Proceedings of the Section on Survey Research Methods*: 421–425.

Williams, R., and Folsom, R. (1981), "Weighted Hot-Deck Imputation of Medical Expenditures Based on a Record Check Subsample." *Proceedings of the Section on Survey Research Methods, American Statistical Association*: 406–411.