Cross-Sectional Imputation and Longitudinal Editing Procedures in the Survey of Income and Program Participation

Prepared for:

U.S. Bureau of the Census

July, 1993

Prepared By:

Steven G. Pennell

The University of Michigan
Survey Research Center
Institute for Social Research
Ann Arbor, Michigan 48106-1248

The author wishes to thank Jim Lepkowski of the Survey Research Center, Angela Feldman-Harkin, Barry Fink, Don Fischer, Donna Riccini, John Coder, Louisa Miller, and Steve Mack at the U. S. Bureau of the Census for their assistance in preparing and reviewing this document, and members of the SIPP Special Interest Working Group for their early comments on the proposed content and organization of the report.

Table of Contents

Table of Tables and Figures

1. Introduction

1.1 Purpose and Audience

This report describes the major cross-sectional imputation and longitudinal editing procedures applied to data collected in the Survey of Income and Program Participation. The report was prepared under a Joint Statistical Agreement between the U.S. Bureau of the Census and the Survey Research Center at the University of Michigan, and is a resource for users of SIPP data products who want an overview of the current cross-sectional imputation and longitudinal editing procedures, or need more details about the imputation and editing procedures to conduct their own evaluation of whether imputed or edited values affect their analyses. The report provides users of SIPP data with an impression of the different types of nonresponse which occur and the nature of the cross-sectional imputations and longitudinal edits performed to compensate for missing or inconsistent data. The audience for this report is the same as that for the "SIPP Users Guide."

Missing data, some of which give rise to the need for cross sectional imputation and longitudinal editing, can be classified on three levels: noncoverage, unit nonresponse and item nonresponse. Noncoverage occurs when units are missing from the sampling frame and therefore are never observed. Unit nonresponse can be described on two levels: complete unit nonresponse in which no interviews are taken within a household; and, partial unit nonresponse in which one or more eligible sample persons, but not all eligible sample persons, within a household are not interviewed. Item nonresponse occurs when interviewed persons are unable or unwilling to provide requested information.

This report describes the SIPP cross-sectional imputation procedures used to compensate for item nonresponse and for selected types of person-level (partial unit) noninterviews. Adjustments for household-level (complete unit) noninterviews are made by increasing the weights of responding persons. Compensation for any noncoverage is generally handled by a post stratification procedure, which is not described in this report. The longitudinal editing procedures in the SIPP are designed to remove inconsistencies in a sample person's longitudinal record introduced through independent wave imputations, to adjudicate occasional disagreements in reported information across waves and to reconcile reported information with Census demographic definitions. The need and desirability of longitudinal editing follows from the practice of processing each wave of SIPP data independently from other waves. The development of the longitudinal file provides an opportunity to incorporate data from surrounding waves into the edit procedures, and in general, to review the record from a longitudinal perspective.

The general meaning of the term "imputation", as used in the report, refers to a general class of procedures which replaces item missing data in one record with nonmissing data from a different record, including statistical matching which is used in the SIPP to compensate for selected types of person-level noninterviews. Changes made to records when they are longitudinally processed are referred to as edits. A distinction is made between cross-sectional imputation and longitudinal editing because in the former case, the procedure is open in the sense that replacement values are acquired

from the records of other sample persons; whereas, in the latter case the procedures are closed because no additional information is obtained from the records of other sample persons.

There are additional aspects of the SIPP data processing procedures which are not covered in this report. Among these related topics are the quality of imputations and the effects imputations and editing have on estimating survey statistics, the development of weights which adjust for household-level noninterviews and the sources and potential biases of nonsampling errors. Many of these related topics are covered in Census publications such as the **SIPP Quality Profile** and reports in the SIPP Working Paper series.

The **SIPP Quality Profile** provides a convenient summary of what is known about the sources and magnitude of errors in estimates based on SIPP data. The profile covers both sampling and nonsampling errors; although, the primary emphasis is on nonsampling errors. The SIPP Working Paper series, some of which are referenced here, cover a wide variety of topics and provide additional reading on nonresponse, imputation and weighting.

1.2 Overview of the SIPP Design

The Survey of Income and Program Participation was initiated in late 1983 (the start of the 1984 panel) by the U.S. Bureau of the Census with the principal objective to provide policy-makers with more accurate and comprehensive information on income and program participation in government programs than were available through other data sources. The survey results were intended to inform policy in the areas of welfare and tax reform, and improvement to entitlement programs such as Social Security and Aid to Families with Dependent Children.

Interviews of panel members by self or proxy reports are conducted every four months for seven or eight consecutive interviews. Each round of interviews is referred to as a wave. Original panel members, also known as 100-level persons, are defined as persons age 15 or older who are living in sampled households on the date of the Wave 1 interview or persons under the age of 15 who become age eligible in subsequent waves. In subsequent waves age eligible persons who join a SIPP sampled household are also interviewed. These persons are known as additional sample persons and are identified by numbers in the 200 plus series, where the leading digit refers to the wave in which the person joined the panel. Each round of interviews collects information on household members for the previous four-month reference period. A new sample or panel is introduced each year. A complete description of the SIPP program is found in Nelson, McMillen and Kasprzyk (1985).

1.3 Organization of the Report

The report is organized into five chapters including the Introduction. Each chapter either describes an imputation or editing procedure applied to information associated with one or more of the SIPP data collection instruments or reviews a related topic.

The basic data collection instruments in the SIPP include the Control Card, the Core questionnaire and one or more Topical Modules. The Control Card is the basic record for each sample unit and contains demographic and household composition information, items transcribed from prior wave interviews as well as administrative data. The Core questionnaire contains questions which are repeated at each interview and are asked of each sample person. Topical Modules contain questions which generally are not repeated at each wave and cover special topics not included in the Core questionnaire.

Despite efforts to ensure a complete set of measures for each sample person in each wave, some persons refuse to be interviewed or cannot be located and other interviewed persons are unwilling or unable to provide all requested information in one or more of the SIPP data collection instruments. In addition, inconsistencies in data between waves only become apparent when a record is viewed longitudinally. In the SIPP, item missing data and selected types on person-level noninterviews are independently imputed in the wave in which the data are missing; i.e., information from preceding or succeeding waves is not used to replace missing information in the current wave during cross-sectional processing. Additional adjustments are made to the data when the records for each wave in a panel are linked together to form the longitudinal file.

Chapter 2 describes the procedure used to impute Core questionnaire items for two types of person-level noninterviews in the SIPP. Chapter 3 outlines the imputation procedures used within a wave to compensate for item missing data in the Core questionnaire. Chapter 3 also notes variations in the Core questionnaire imputation procedures used to compensate for item missing data in the Control Card and Topical Modules. Chapter 4 provides an overview of the longitudinal editing procedures. Chapter 5 reviews strategies for assessing the influence of imputed data on one's analyses.

1.4 Types of Noninterviews and Item Missing Data in the SIPP

The U.S. Bureau of the Census classifies noninterviews at both the household and person level. Household level noninterviews occur when a housing unit is sampled but no interviews are obtained because the housing unit not longer exists, is not occupied, cannot be located or the occupants of the household are temporarily away or refuse to be interviewed. Person-level noninterviews are defined only in households in which at least one person was interviewed and occur because one or more sample persons, but not all sample persons in the household, refuse to be interviewed or are unavailable and a proxy report is not obtained. Person-level noninterviews may occur for one wave, two or more consecutive waves or between responding waves. The various types of household and person noninterviews are reviewed below.

Household-Level Noninterviews

Type A Noninterview: Type A noninterviews consist of households occupied by persons eligible for interview and for whom a questionnaire would have been completed if an interview had been obtained. Type A noninterviews occur when every eligible member of the household is a noninterview. Type A noninterviews occur when no one is at home in spite of repeated visits, all household members are temporarily absent during the entire interview period (for example, they are away on vacation), household members refuse to participate in the survey, the household cannot be located, the housing unit cannot be reached because of impassable roads or interviews cannot be taken because of serious illness or death in the family.

Type B Noninterview: this type of noninterview occurs when a housing unit is vacant, occupied by persons with their usual residence elsewhere, unfit for occupancy or set to be demolished, under construction and not ready for occupancy, or converted to temporary business or storage. It also occurs when a site for a mobile home, trailer or tent is unoccupied or when a permit has been granted, but construction has not started.

Type C Noninterview: occurs when a housing unit is demolished, or house or trailer is moved, converted to permanent business or storage, merged or condemned. These later reasons apply in Wave 1 only. In subsequent waves, Type C noninterviews are defined when **all** sample persons are deceased, have moved outside the country or are living in armed forces barracks.

Type D Noninterview: Type D noninterviews only occur in Waves 2 and beyond and are defined when a household or some members of a household are living at an unknown new address or at an address located more than 100 miles from a SIPP sample area and a telephone interview is not conducted.

Person-Level Noninterviews

Type Z Noninterview: Type Z noninterviews occur when a member of an interviewed household is not interviewed because they are unavailable for an interview or refuse and a proxy interview is not obtained.

Departure Noninterview: is defined by someone who was a member of a SIPP interviewed household sometime during the four-month reference period but was no longer a household member on the date of interview. The phrase "Departure Noninterview", which is not an official Census term, is used throughout this report as a convenient way to distinguish between the two types of person-level noninterviews.

Item Nonresponse

Item Nonresponse: item nonresponse occurs when a response to one or more questions is not provided, though most of the questionnaire is completed.

Among these four types of household noninterviews, no adjustment is required to compensate for Type B and Type C noninterviews. This is because Type C noninterviews are no longer housing units at the original sample address or the housing unit no longer is occupied by sample persons. Housing units classified as Type B noninterviews either have no occupants or the occupants' usual residence is elsewhere. Persons whose usual residence is elsewhere are not interviewed because they have a chance of being in the sample at their usual residence. Weighting adjustments are used to compensate for Type A and Type D noninterviews. Item nonresponse and both types of person-level noninterviews in the SIPP are imputed using the procedures described in this report.

1.5 Sequence of SIPP Cross-Sectional Imputations and Longitudinal Edits

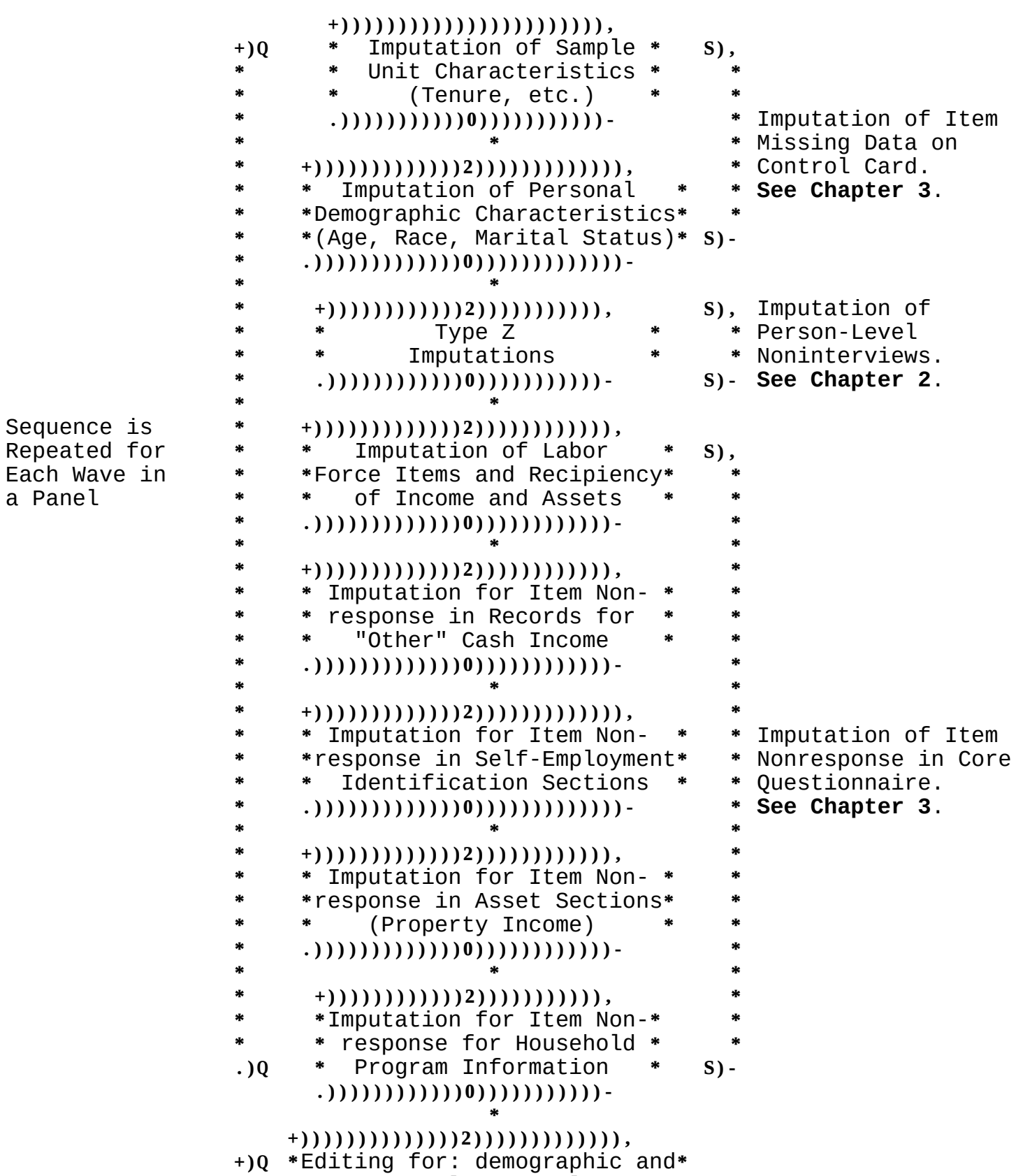
Figure 1.1 outlines the sequence of steps in which the SIPP data are processed cross sectionally and longitudinally. The presentation of material in this report generally follows the sequence of processing steps in Figure 1.1. When SIPP data are processed cross sectionally each wave of data are treated separately. The cross-sectional processing begins by imputing item missing data on the Control Card.[1] Missing items on the Control Card are imputed first because many of the demographic variables located there are used in subsequent imputation steps and need to be nonmissing for all cases.[2] Next, Core questionnaire records are imputed in full from a single donor for two types of person-level noninterviews. Because person-level noninterviews are imputed before donor records are processed for item missing data, imputed noninterview records initially retain the pattern of item missing data on the donor record. Missing items on the Core questionnaire are subsequently imputed for responding sample persons and for noninterviews whose records were previously imputed. The processing of Core questionnaire items is also sequenced so that missing items in earlier steps can be used to impute missing items in later steps.

---

[1] At the same time item missing data are imputed a series of logical consistency and other edit checks are also applied to the data.

[2] The imputation of item missing data in the Control Card is covered in Chapter 3, which describes the imputation of item missing in the Core questionnaire.

Figure 1.1 **Sequence of Cross-Sectional Imputation and Longitudinal Editing Procedures (Adapted from Nelson, McMillen and Kasprzyk, 1985)**



See Chapter 4.

Item missing data on the Core questionnaire are imputed section by section in the following sequence:

1. Labor force and recipiency;
2. Other cash income;
3. Wage and salary and self-employment variables;
4. Asset variables; and
5. Program participation variables.

Item missing data on Topical Modules are imputed at the same time missing items on the Core questionnaire are imputed. Once the data for each wave in a panel has been processed, selected groups of items are extracted from each wave and longitudinally edited. The process of extracting and editing is performed separately for the following groups of items:

1. Demographic and household variables;
2. Employment variables;
3. General amount variables; and
4. Other variables.

As each group of items is edited they are joined together to create the SIPP Longitudinal file.

1.6 Goals of Imputation

There are two general goals of imputation, one is statistical and the other is practical. The statistical goal of imputation is to minimize the mean square error of survey estimates. The mean square error has both a variance and a bias component. All imputation procedures increase the variance of estimates but some imputation procedures increase the variance less than others. Imputation can reduce the bias component of the mean square error to the extent systematic patterns of item nonresponse are identified and correctly modeled. The ability of an imputation scheme to correctly guess the missing values of individual items is of lesser importance; although, the better an imputation scheme is able to does this, the smaller will be the error due to imputation. The statistical goal of imputing missing data in the SIPP is also more general than specific. The SIPP imputation procedures are not designed to address estimation of specific parameters, but rather to provide reasonable estimates for a variety of analytical purposes. No single imputation procedure is likely to be ideal for all analytical purposes.

There are also several practical goals for imputing missing data. Consistency is maintained between the results from different analyses when missing data are imputed because cases with missing data are not necessarily excluded. In the absence of imputation, and in the presence of missing data, different analyses will be based on different subsets of cases depending on the pattern of missing data. For analyses based on casewise deletion of missing data, partial information about otherwise responding cases is sacrificed. The construction of household and family level variables is also made easier when missing items on individual records are imputed.

Although the statistical goal of imputation is to reduce the bias component of the mean square error, there is no guarantee that estimates based on imputed data are less biased than estimates based only on nonmissing data. In fact, the bias associated with estimates based on imputed data could be greater depending on the type of imputation used and the parameter being estimated. Imputation also has the distinct disadvantage of creating the impression that the data are complete. All statistical imputation procedures fabricate data which increase the variance of estimates. Because the increase in variance due to imputation is difficult to incorporate into variance estimates, the precision of survey estimates is often overstated. In essence, imputation can reduce the effective sample size of the data file.

2. Compensation for Person Noninterviews: Type Z Imputation of Core Questionnaire

2.1 Introduction

This chapter describes a statistical matching procedure used to impute wave nonresponse for two types of person-level noninterviews in otherwise cooperating SIPP households.[3] The procedure imputes an entire Core questionnaire from a single donor for both types of noninterviews when less than the full complement of eligible sample persons is interviewed in a household. The first type of imputed noninterview is for persons aged 15 or older who were members of interviewed households at the beginning of the four-month reference period but were not members of any SIPP interviewed household on the date of interview. These persons may have moved within the United States and were not located, or moved outside the United States, entered institutions, moved to armed forces barracks, died or became ineligible for interview because they were no longer living with a 100-level sample person. Throughout this report persons who were members of SIPP interviewed households at the beginning of the reference period but not on the date of interview are referred to as "Departure" noninterviews. Although the term "Departure" noninterview is not an official Census phrase, it is used in the report as a convenient way to distinguish between the two types of imputed noninterviews.

The second type of imputed noninterview is for persons aged 15 or older who were members of SIPP interviewed households on the date of interview and during all or a portion of the four-month reference period, but were not interviewed because they refused to cooperate or were unavailable for interview and a proxy report was not possible. These sample persons are referred to as "Type Z" noninterviews. The letter "Z" is used to distinguish person-level noninterviews from the various types of household-level noninterviews--Types A, B, C and D.

---

[3] Weighting adjustments are used to compensate for Type A and D noninterviews. Item missing data are handled by a sequential hot-deck imputation procedure which is described in Chapter 3. See Chapter 4 for a description of longitudinal editing procedures.

2.2 Reasons for Imputing Wave Nonresponse

Core questionnaires for both types of noninterviews are imputed so that information reported by other household members can be retained and aggregate household variables such as income can be constructed. Moreover, by imputing from a single donor rather than from multiple donors, the interrelationships between variables are preserved. The alternative is to consider the entire household as a noninterview and to compensate for the noninterview by weighting households in which all eligible persons responded.

Both types of noninterviews are imputed for each wave in which the sample person was eligible, but not interviewed. Type Z noninterviews can have imputed data for one wave and self or proxy reported data for the surrounding waves, or they can have imputed data for multiple waves, some of which can be consecutive.[4] Departure noninterviews, however, will most likely have imputed data for only one wave, and, with few exceptions, will not reappear in subsequent waves, unless they rejoin a SIPP interviewed household from an institution or service in the Armed Forces. Persons who leave an eligible sample household sometime during the Wave 1 reference period are never classified as Departure noninterviews because the SIPP sample is defined by eligible persons residing within a selected household on the date of interview.

Type Z noninterviews and Departure noninterviews can include:

* Original sample persons first interviewed in Wave 1 (100-level persons) but who are not interviewed in one or more of the following waves for which they remain eligible;

* Sample persons who become age eligible (turn 15) after Wave 1 but who are not interviewed in one or more of the following waves for which they remain eligible; and

* Persons who join the household of an original sample person after Wave 1 (referred to as additional persons, nonsample persons or 200 plus persons) but who are not interviewed in one or more of the following waves for which they remain eligible.

---

[4] Persons who were members of SIPP interviewed households at the beginning of the reference period but not on the date of interview are identified on individual wave files as follows: POP-STAT=1 (person was age 15 or older in month of interview) and PP-MIS*=1 for *=1 or *=1 and 2, or *=1, 2 and 3; or *=1, 2, 3 and 4 (person was in sample for at least one month of the reference period starting in month 1) and PP-INTVW=0 (person was not interviewed). These persons are identified on the Full Panel Research File as follows: AGE(CMONTH) is 15 or more, where CMONTH=weighting control month and PP-INTVW=0 and where PP-MIS=1 for month one of the corresponding wave for one or more waves in the period covered by the weight, i.e., calendar year 1, calendar year 2, panel weight.

Type Z noninterviews are identified on individual wave files as follows: PP-INTVW=3 OR 4.

2.3 Examples of Type Z and Departure Noninterviews

Table 2.1 provides examples of household membership patterns across the four-month reference period preceding the month of interview as well as the month of interview and indicates whether data for person-level noninterviews with that pattern are imputed. Persons with patterns 1, 2 or 3 in Table 2.1 represent Type Z noninterviews and are imputed in each wave they are a noninterview. Noninterviews with pattern 3 are considered Type Z noninterviews because sample persons entered SIPP interviewed households on or before the 15th of the interview month. All persons age 15 or older who enter SIPP interviewed households on or before the 15th of the interview month are considered eligible household members whose Core questionnaire data are imputed if not interviewed. Persons who enter SIPP interviewed households after the 15th of the interview month (pattern 4), on the other hand, are not considered eligible household members and do not appear in the cross-sectional file because they are classified as out of sample for all months of the wave. Noninterviews with household membership patterns 5, 6 or 7 are considered Departure noninterviews whose Core questionnaire data are imputed. Pattern 7 represents sample persons who leave a SIPP interviewed household after the 15 of the last reference month of the four-month reference period. These persons are classified as eligible household members whose data are imputed. Persons who leave SIPP interviewed households on or before the 15 of the last reference month (pattern 8) are ineligible because they are defined as out of sample for the entire four-month reference period. Persons who both enter and leave a SIPP interviewed household during the reference period and are gone before the month of interview, such as those indicated by pattern 9, are never recorded as household members; consequently, no imputation is performed for persons with this pattern of household membership.

2.4 Type Z Imputation Procedures

The methods used to impute records for noninterviews with household membership patterns 1, 2 or 3 (Type Z noninterviews) and 5, 6 or 7 (Departure noninterviews) are called Type Z imputation procedures.[5] Type Z imputation is based on a hierarchical sorting and merging operation which matches noninterviews with respondents on socioeconomic characteristics available for both. Type Z imputation procedures are designed such that a match is always found. Once a matching donor is identified Core questionnaire values reported by the donor, or provided by a proxy, are assigned to the noninterview record in full, except for identification variables or other variables not relevant for the household in which the noninterview occurred.

[5] Note: Topical Module data for Type Z and Departure noninterviews are imputed using the item-by-item sequential hot-deck procedure employed for item missing data in the Core questionnaire. See Chapter 3 for an overview of the SIPP sequential hot-deck procedure.

Core questionnaire data for Type Z and Departure noninterviews are imputed in the following three steps.

1. In the first step, noninterview (recipient) and respondent (donor) cases are identified.

2. In the second step, each noninterview case is matched with four or five donor records depending on which matching variables are available and whether sample persons were interviewed in the previous wave.

3. One donor record which represents the "best" match is selected in the final step.

The donor pool from which a respondent record is selected and duplicated for a noninterview case includes all persons age 15 or older who were interviewed or whose information was collected by proxy in the wave in which the noninterview occurred. The universe of donors for a particular wave is always restricted to respondent records in that wave; respondent records from preceding or succeeding waves are never used, although some matching variables may be obtained from the preceding wave.[6] The identification of donor records or whether the donated data were obtained by self or proxy reports is not indicated on the recipient record.

2.4.1 Matching Variables

The socioeconomic variables which are used to match noninterviews with respondents are either taken from the current wave Control Card, extrapolated forward from previous wave Control Card information or, if missing on the Control Card, imputed for the current wave using an item by item hot-deck procedure (see section 3.5.8). The variables used to match noninterviews with respondents include age, race, gender, marital status, household relationship, education, veteran status, parent/guardian status and income and asset sources. Age and marital status are defined at two levels, with one level containing more detail than the other. Figure 2.1 provides a description of the global set of 10 matching variables, subsets of which are used to match noninterview and respondent records on several levels.

[6] The pool of eligible donors is identified on the individual wave files as follows: AGE(5) GE 15 and PP-INTVW EQ 1 OR 2.

Table 2.1 Examples of Household Membership Patterns by Reference Month and Month of Interview for Person-Level Noninterviews and Whether the Noninterview is Imputed*

	Type Z Noninterview	Departure Noninterview	Noninterview Imputed	Reference Month				Interview Month
				4	3	2	1	0
1	Yes	No	Yes					
2	Yes	No	Yes					
3	Yes	No	Yes	(enter on or before 15)				
4	No	No	No	(enter after 15)				
5	No	Yes	Yes					
6	No	Yes	Yes					
7	No	Yes	Yes				(leave after 15)	
8	No	No	No				(leave on or before 15)	
9	No	No	No					

* Shaded areas in Table 2.1 indicate months noninterviewed person was a member of a SIPP interviewed household.

Income and asset variables which are used to match noninterviews with respondents are obtained from the previous wave Control Card for both current wave noninterviews and respondents if an interview was obtained in the previous wave; otherwise, income and asset variables are not used as matching variables. Consequently, income and asset variables are never used as matching variables to impute Wave 1 noninterviews. All other matching variables are obtained from the current wave Control Card, which may contain items extrapolated from the previous wave Control Card or imputed, as noted above.

Income and asset matching variables are obtained from the previous wave Control Card when available for both donors and recipients for two reasons. First, the current wave is unlikely to contain income and asset information for noninterviews. Second, if income and asset variables for noninterviews are obtained from the previous wave then corresponding variables for respondents are also obtained from the previous wave to maintain temporal consistency. These procedures increase the likelihood that a better quality imputation is achieved because additional information about a noninterview is used when available. If the matching variables are correlated with missing values then as the matching groups become more homogeneous, which is a function of the number of matching variables, any nonresponse bias should be reduced.

Figure 2.1 Variables Used to Match Recipients with Donors for Imputing Two Types of Noninterviews in the SIPP

AGE (AGE1)
1. 15 to 17
2. 18 to 24
3. 25 to 54
4. 55 to 64
5. 65 or older

AGE (AGE2)
1. 15 TO 24
2. 25 TO 54
3. 55 to 64
4. 65 or older

RACE
1. White and other nonblack
2. Black

SEX
1. Male
2. Female

MARITAL STATUS (MS1)
1. Married [MS(5)=1,2]
2. Divorced or separated [MS(5)=4,5]
3. Widowed [MS(5)=3]
4. Never married [MS(5)=6]

MARITAL STATUS (MS2)
1. Married [MS(5)=1,2]
2. All others [MS(5)=3,4,5,6]

HOUSEHOLD RELATIONSHIP
1. Householder [RRP(5)=1,2]
2. Not householder [RRP(5) NE 1,2]

EDUCATION ATTAINMENT
1. Less than 16 years completed
2. 16 or more years completed (HIGRADE GT 24 OR (HIGRADE EQ 24 AND GRD-CMPL EQ 1))

VETERAN STATUS
1. Veteran (U-VET EQ 1 AND U-AF NE 1)
2. Nonveteran (ALL OTHER)

DESIGNATED PARENT/GUARDIAN
1. Yes
2. No

INCOME SOURCES (See Appendix 1 for a description of income sources)
1. Had no income sources 1-56
2. Had income sources 1 or 2
3. Had income sources 3 or 20-24
4. Had source 27
5. All other combinations

ASSET SOURCES (See Appendix 1 for a description of asset sources)
1. Had no asset sources 100-150
2. Had asset sources 105-107 or 120-150
3. Had sources 100-104 or 110

2.4.2 Matching Levels

In practice, a noninterview cannot always be matched with a respondent on all matching variables. To account for situations where a match cannot be made on all variables, simultaneous matches are made at several lower levels of detail by omitting some matching variables and reducing the number of categories in others. The socioeconomic variables which define each match level are outlined in Table 2.2. In total, there are 9 different match levels which are organized into two groups depending on the interview status of the sample person in the previous wave. The two groups of matching levels are:

1. Persons interviewed in the previous wave are matched at five levels referred to as Type B matches; and
2. Persons not interviewed in the previous wave are matched at four levels referred to as Type A matches.

Table 2.2 Matching Levels Used to Impute Core Questionnaire for Two Types of Noninterviews in the SIPP

Matching Variables	Type B Matching Levels for Persons Interviewed in Previous Wave					Type A Matching Levels for Persons not Interviewed in Previous Wave			
	1*	2	3	4	5	1	2	3	4
Age (AGE1)	■	■	■	■		■	■	■	
Age (AGE2)					■				■
Race	■	■				■	■		
Sex	■	■	■	■	■	■	■	■	■
Marital Status (MS1)	■					■	■		
Marital Status (MS2)		■	■	■	■			■	■
Householder Status	■	■	■			■	■		
Education	■	■	■			■	■	■	
Veteran Status						■			
Parent/Guardian	■	■	■	■	■	■	■	■	■
Income Types	■	■	■	■					
Assets Sources	■	■							

* Shaded areas indicate variables which define level of match.

Each matching level is defined by a different combination of variables. No level is defined by all 10 matching variables. Age, gender and parent/guardian status are common to all match levels, and income and asset variables are used only if a sample person was interviewed in the previous wave. The matching levels are arranged hierarchically and become progressively less demanding on the qualifications for a match. The least demanding level is defined such that a match is always found at that level.

A noninterview case is matched at several levels against a pool of donors to avoid the potential problem of not finding a match when only a large number of matching variables are used, which produces a potentially better quality imputation, and getting an imputation of potentially lesser quality when only a few matching variables are used. Matching noninterviews with respondents at several levels ensures that a donor record is always found, and if a match is found at more than one level, the one "best" match is selected.

2.4.3 Matching Operation

Once current wave noninterviews and donors are identified and their interview status for the prior wave is determined two files are created : one file is created for noninterviews and one file is created for respondents, each of which contains a number of matching records. The noninterview file contains either 5 Type B records or 4 Type A records for each noninterview case depending on whether the sample person was interviewed in the previous wave. Each record created corresponds to one of the matching levels outlined in Table 2.2. Comparable records are created for sample persons who qualify as donors: if the donor was interviewed in the previous wave, 5 Type B **and** 4 Type A records are created for each donor case; otherwise only 4 Type A records are created because the donor was not interviewed in the previous wave. Each record created has information about the sample unit and sample person and contains three match keys:

1. Type of match, A or B;
2. Level of match, 1-4 or 1-5 depending on type of match; 3. A match index which represents the product of the values of each of the match variables for each type and level of match.

The two files are subsequently sorted by and compared on match type, match level and match index.

The output from the matching operation is an updated file of noninterview matching records. When a match for a particular level is found the noninterview matching record is updated with information from the respondent matching record. If no match is found, the noninterview matching record is updated with zeros, indicating no match was found for the noninterview at that level. The donor record selected and duplicated for the noninterview corresponds to the level at which the best match was found. The best match occurs at the level containing the most variables with the greatest

amount of detail. The best match always corresponds to the lowest numbered level at which a match occurred.[7]

Donor selection is sequential with replacement, i.e., donors are selected within a matching group from top to bottom. A matching group is defined by records having the same value on the match index. When there are more noninterviews than donors in a matching group the donor pool is recycled from top to bottom. In this way a respondent may serve as a donor more than once. Appendix 2 provides an overview and illustration of the Type Z imputation procedures.[8]

Type Z imputations are performed before respondent records have been edited; although, some items on the respondent record which would be inappropriate for the noninterview case are removed before the imputation procedure begins. For example, the respondent's health insurance policy may contain information about covered dependents which may not be relevant or appropriate for a noninterview case and is removed from the respondent record prior to imputation. References to "person numbers" are also removed from the respondent donor record wherever they occur prior to imputation. Since donor records can contain item missing data or other items considered to be inconsistent prior to being edited, the imputed records initially contain any missing and inconsistent items present on the donor records. Type Z-imputed records and interviewed records are subsequently edited and imputed for item missing data together. Consequently, once all records have been edited, a Type-Z imputed record and its donor may no longer have the same set of measurements. The extent to which this is true depends on the level of missing data on the donor record and the changes made during the editing process.

These procedures which impute Core questionnaire data from a single donor for both Type Z and Departure noninterviews are applied consistently within waves and across panels with one exception: Type Z noninterviews were not imputed for Wave 1 of the 1984 Panel.[9] Data for both types of noninterviews are imputed for each month in the reference period whether or not the person was in sample for a particular month. Zero weights are assigned to reference months in the wave files in which an imputed noninterview was not in sample. In the longitudinal files, all monthly fields for reference months with PP-MIS=0,2 (person was in a Type A noninterview household or was not in sample for that month) are dropped from the record.

---

[7] The type of match, level of match and whether the donor information was self reported or obtained through a proxy are not carried on the wave or longitudinal files for imputed cases.

[8] See Coder (1978) and Welniak and Coder (1980) for additional reading about this hierarchical imputation procedure.

[9] Type Z noninterviews in Wave 1 of the 1984 panel were classified as Type A noninterviews, and compensated by weighting respondent cases, because an imputation system was not fully developed at the time.

Table 2.3 Number of Type Z-Imputed Records for Selected Waves and Panels by Type and Level of Match

		Type A: Level of Match				Type B: Level Of Match					
Panel	Wave	1	2	3	4	1	2	3	4	5	Total
1990	1	657	3	1	0	0	0	0	0	0	661
	2	699	1	2	0	1034	17	10	3	0	1766
	3	1093	1	4	0	986	11	7	6	4	2112
	4	1278	3	3	0	1148	11	13	5	6	2467
	5	1575	1	5	1	1179	10	27	3	2	2803
	6	1737	0	7	2	942	8	4	2	3	2705
	7	1788	2	5	1	875	11	13	5	3	2703
	8	1858	3	4	1	883	10	7	3	0	2769
1991	1	360	0	3	0	0	0	0	0	0	363
	2	421	1	1	0	670	11	16	5	1	1126
	3	754	2	7	3	654	7	8	8	2	1445
	4	876	2	6	0	602	5	14	2	0	1507
	5	1012	1	2	1	634	10	8	6	0	1674
1992	1	819	0	5	1	0	0	0	0	0	825

2.4.4 Number of Type Z Imputed Records for Selected Panels and Waves

Table 2.3 contains the number of Type Z-imputed records by type of match (A or B) and level of match (1-4 for Type A level matches and 1-5 for Type B level matches) for selected panels and waves. The total number of Type Z-imputed records for a particular wave is also noted. The distribution of imputations by level of match shows that the largest number of imputations occur at level 1, which corresponds to the level containing the most matching characteristics of any level. By definition there are no Type B matches for wave 1 of a panel.

3. Compensation for Item Missing Data in the Core Questionnaire, Control Card and Topical Modules

3.1 Introduction

This chapter describes the imputation procedures used in the SIPP to compensate for item missing data in the Core questionnaire and Topical Modules for responding sample persons and Type-Z imputed noninterviews, and variations applied to the procedure for imputing item missing data in the Control Card.[10]

In general, imputation of item missing data is the process of replacing missing or inconsistent data in one case with plausible values selected from another case. Item missing data occur when respondents refuse to provide or do not have the information requested, or the information provided is inconsistent with edit specifications and the response is deleted during the processing stage; or arise when interviewers forget to ask for the information, or record it incorrectly and an edit failure results; or are introduced as keying errors during the processing stage. In the SIPP plausible values for item missing data are identified by logical deduction or by statistical imputation. Each of these procedures is described in this chapter.

3.2 Reasons for Imputing Item Missing Data

The main objective of imputing item missing data is to reduce the mean square error of estimates, particularly the nonresponse bias component which arises when the underlying process which generates the pattern of item missing data is not random.[11] Data sets which are imputed for item missing data offer the convenience of being easier to analyze because data users can perform a wide range of statistical analyses without omitting cases with missing values on one or more variables. Moreover, carefully designed imputation procedures preserve interrelationships among variables and provide consistency between results from different analyses. This last feature is important for public use data sets such as the SIPP where the different analyses to be performed on the data are not known a priori.

The alternative to imputing item missing data is to do nothing. But doing nothing implies a model that item missing data are missing at random among all sampled cases, an unlikely assumption at best which can lead to biased estimates. Although all imputation procedures are based on models, the models are often more explicit and plausible. For example, instead of a model that assumes item missing data are randomly distributed among all cases, the most widely used imputation procedures assume the more realistic model that item missing data are randomly distributed within subgroups.

---

[10] See Section 2 for a discussion of the imputation procedures used to compensate for unit nonresponse within otherwise responding households and Section 4 for a discussion of the longitudinal edits and imputations.

[11] For further discussion of the mean square error see Kish (1965).

3.3 Components of an Item Missing Data Imputation Procedure

There are numerous components of any item missing data imputation procedure and variations in how they are configured in practice. For instance, the source of the replacement values may be the current data set, other sample surveys, censuses or administrative records. The type of replacement values may be individual attributes or quantities, means, model-based predictions, values based on a distance function or expert judgement. The imputations may be carried out independently within subclasses of the population. Each missing item may be imputed independently or groups of related items may be imputed together, receiving replacement values from the same donor. The procedure for selecting donors may be random, implemented with or without replacement, nonrandom or sequential from an ordered file. Moreover, not all respondent cases may be eligible to serve as donors. For example, restrictions may be placed on respondent cases with outlying values or values which were obtained by proxy reports rather than from the donor directly. Panel surveys such as the SIPP have the additional option of imputing for a missing response to an item in one wave with the nonmissing response to the same item provided in another wave.

The way in which these components are configured in practice consider the manner in which the data will be used and resource constraints such as time and money. Some imputation procedures are appropriate for data collected solely for producing descriptive measures, while other imputation procedures are more appropriate for data to be used analytically. Some imputation methods alter marginal distributions and distort variance-covariance matrices as well. For large ongoing surveys like the SIPP imputation systems need to be set up ahead of time and implemented quickly and efficiently on an ongoing basis. Moreover, the imputation system needs to anticipate a variety of analytical applications to be made with the data.

The following sections describe the two imputation methods used to compensate for item missing data in the SIPP Core questionnaire and Topical Modules: logical, or deductive, imputation and a statistical imputation procedure known as a sequential hot-deck, and variations in the sequential hot-deck procedure applied to missing items on the Control Card.

3.4 Logical Imputation of Item Missing Data

Logical, or deductive, imputations are preferred over statistical imputations and are used in the SIPP whenever missing or inconsistent items can be reasonably inferred from nonmissing items within the same record. The advantage of logical imputation is that the increase in variance due to imputing several missing items on one record with nonmissing items from numerous other records is avoided. Logical imputations are used to replace missing items in the Core and Topical Module questionnaires and in the Control Card. A general overview of logical imputation in the SIPP is provided below, but details of the large class of edits which encompass logical imputation are not within the scope of this report.

Prior to the point in the cross-sectional editing process at which a missing item would be imputed a check is made for feasible values within the section the missing item is located. If a feasible

value can be inferred from reported information, the inferred value replaces the missing value and no statistical imputation of that item for that case is performed. For example, when an answer to the question about looking for work during the reference period is missing, a "yes" answer is logically imputed if the respondent indicated in subsequent items that they looked for work during one or more weeks in the reference period. If appropriate, the search for feasible responses is sought among nonmissing items both before and after the missing item. There are no imputation flags in the SIPP cross-sectional or longitudinal files which indicate that the value of an item was logically imputed, nor is a distinction made between data obtained by self or proxy interviews.

3.5 Sequential Hot-Deck Imputation Procedure in the SIPP

The statistical imputation procedure used in the SIPP to compensate for item missing data in the Core questionnaire, Control Card and Topical Modules is referred to as a sequential hot-deck. Although there is no general agreement on the precise definition of a hot-deck procedure, in the SIPP it refers to an imputation procedure which replaces item missing data in one wave with nonmissing values from different interviewed cases in the same wave. The main advantage of the hot-deck method for survey programs such as the SIPP which collect large amounts of data is that it generally produces feasible values because the replacement values are taken from the same wave in which items are missing. Selecting replacement values from the same wave in which items are missing improves the likelihood that the distributional properties of the sample are maintained. The hot-deck method may not be ideal for imputing missing data in surveys with smaller samples or where analytical requirements are very specialized.

Procedures which impute an overall or stratum mean from the current data set are not hot-deck methods according to this definition. A possible disadvantage of the mean value imputation method is that it distorts the distributional properties of the sample. "Cold-deck" imputation procedures, in comparison, are based on methods which select values or use relationships obtained from sources other than the current data set.[12] Cold-deck imputation procedures do not always provide feasible values and are rarely used today except to provide starting values for a hot-deck procedure.

The hot-deck procedure used in the SIPP is sequential because the selection of replacement values is implemented one record at a time from an ordered file. The sequential nature of the procedure, however, may also be its greatest disadvantage. Because the order of the file is not based on a probability mechanism, a model-free theoretical evaluation of the procedure is not possible. Also, the procedure may give rise to multiple uses of the same donor (Kalton, 1983).

The sequential hot-deck procedure used in the SIPP is carried out independently for each wave and by groups of related variables within the Core questionnaire, and involves five key

---

[12] For an overview of various imputation procedures see, for example: Kalton and Kasprzyk (1982, 1986) or Sande (1982, 1983).

steps:

1. Specifying cold-deck or starting values;

2. Sorting the sample cases;

3. Preprocessing the data file to identify records with no item missing data and to update cold-deck values.;

4. Classifying cases into subclasses of the population, referred to as imputation classes or adjustment cells, according to values on a set of classification or auxiliary variables which are nonmissing for all cases; and

5. Selecting replacement values from donor cases to impute item missing data on recipient cases.

The groups of related variables processed separately are:

* Labor force and recipiency of income and asset items;

* Other cash income items;

* Wage and salary and self employment variables;

* Asset income items; and

* Program information variables.

An example of the SIPP sequential hot-deck procedure is provided in Section 3.5.6. The set of classification variables used to impute item missing data in the 1987 Panel Core Questionnaire is contained in Figure 3.2 (page 38) and discussed in Section 3.5.10. The various imputation matrices are also described in Section 3.5.10 and defined in Tables 3.1 through 3.4 (pages 44-49). Figure 3.3 (page 50) and Figure 3.4 (page 53) show the correspondence between Core Questionnaire variables and the imputation matrices, as well as starting cold-deck values. Table 3.5 (page 54) contains rates of item missing data for selected variables, panels and waves.

3.5.1 Specifying Starting or Cold-Deck Values

The sequential hot-deck imputation procedure used in the SIPP begins by filling a large matrix with starting or cold-deck values. The cells in the matrix are defined by the cross classification of auxiliary variables. Each cell in the matrix corresponds to respondent cases with the same set of values on the classification variables. Many different matrices are defined in the SIPP and each matrix corresponds to one or more variables subject to imputation.

The matrix is initially referred to as the cold-deck matrix. During subsequent stages of processing, as the cold-deck values are replaced with information from the current wave, the array of cells is referred to as the hot-deck matrix. Historically, cold-deck values in a sequential hot-deck procedure served as the initial set of replacement values for missing items in the first record processed; missing items in subsequent records would typically receive replacement (hot-deck) values from the current data set. In the SIPP, however, cold-deck values are not frequently used as replacement values for either the first or subsequent records processed. The primary purpose of cold-deck values in the SIPP is to initialize the cold-deck matrix.

The cold-deck values which initially fill the matrix are specified by subject matter specialists at the Bureau of the Census and come from previous SIPP surveys, administrative records, other surveys or censuses. Starting cold-deck values for Core questionnaire items generally do not change within a panel. Changes in starting cold-deck values between panels, however, are more common. Starting cold-deck values for items in the Topical Modules, in comparison, change more frequently because they generally are more sensitive to changes in economic activity.

3.5.2 Sorting the Sample Cases

The records in the sample file are sorted by three geographic variables prior to imputing item missing data. The three geographic sort variables are primary sampling unit, segment number and serial number. The cases are sorted prior to processing and are not resorted at any other time during the imputation process. The sorting operation creates a file in which neighboring records represent geographically proximate households.

3.5.3 Preprocessing the Sample File: Initial Updating of Cold-Deck Values

Once the cases have been sorted they are processed through a series of edit programs. During the first pass against the edit programs the cold-deck values in the matrix are updated with information from the current wave, but missing data are not imputed. The imputations are performed during the second pass through the edit programs. The initial processing is done separately (but simultaneously) for each group of related Core questionnaire variables outlined above.

During the first pass against the edit programs the first record in the sorted file with consistent and nonmissing data for a particular section is identified and the values from this case replace the cold-deck values for that section in the matrix. The values for each subsequent record with consistent and nonmissing information in a section update the previous set of consistent and nonmissing values written to the matrix. The initial updating procedure is performed case by case rather than item by

item to insure that the initial set of replacement items for a particular section are consistent. The checking and updating operation continues until all the records in the data file have been processed. The last set of values written to the matrix serve as the starting values in the subsequent sequential hot-deck procedure. In this way, cold-deck values are rarely used as replacement values in the SIPP because the initial processing usually replaces all starting values with values from the current wave of data collection. The initial set of hot-deck values obtained in this way probably represent a number of different responding cases across sections due to the pattern of item missing data.

3.5.4 Allocating Cases into Imputation Classes

In the next step of the imputation procedure each respondent and noninterview record in the sorted file is allocated to one of the imputation classes or adjustment cells according to its values on the set of classification or auxiliary variables. Each matrix is defined by a different set of classification variables and corresponds to a single item or to a series of related items whose missing data are imputed. The number of cells in each matrix is equal to the product of the number of levels in each classification variable. An imputation matrix defined by 5 variables, each measured on a three point scale, for example, has a total of 243 cells. More than 50 different sets of classification variables are used in the SIPP to impute item missing data in the Core questionnaire and the number of cells in these matrices range from under 100 to over 1,000. The sets of classification variables are fully described in Section 3.5.10 and in Figures 3.2 and 3.3 and in Tables 3.1 through 3.4.

The selection of classification variables is determined by the subject matter specialists at the Bureau of the Census who base their selections on the extent to which the nonmissing values of the variable being imputed are correlated with the classification variables, the extent to which the classification variables are nonmissing for all cases and the linkages through edits. Ideally, the set of classification variables should account for a large proportion of the variance in the variable being imputed and be associated with variations in response rates.

The allocation of sample cases into imputation classes (also known as subclasses or strata) according to a set of classification variables serves several purposes. The classification procedure creates homogeneous adjustment cells such that cases within an adjustment cell are more similar than cases between adjustment cells. In this way donors and recipients are similar under the assumption that the nonresponse mechanism within the imputation class is not related to the item being imputed; that is, an underlying assumption is made that item missing data are distributed randomly within the subclass defined by the cross classification of the auxiliary variables. The selection of classification variables may also place bounds on the range of values which can be imputed and implicitly satisfy edit constraints. The implicit stratification created by the sort order of the file further improves the opportunity for a better quality imputation to the extent that the sort order creates positive autocorrelation, where nearby cases are more similar to each other than cases which are further apart in the file.

3.5.5 Subsequent Updating of Hot-Deck Values and Replacement of Missing Values

The selection of replacement values for missing items is restricted to donor and recipient records within each particular cell; that is, nonmissing records allocated to one cell never denote information to records with missing items in another cell. As the file is processed through the set of edit programs the second time and the imputations are performed, the set of hot-deck values is updated once again, but this time the updating procedure is item by item rather than case by case. Missing items in the first record processed receive the final set of replacement values which were obtained from the initial updating procedure. The nonmissing values in the first record processed update the corresponding set of current hot-deck values. These current hot-deck values, in turn, donate information to any missing items in the next record processed.

The records are processed one at a time in a sequential fashion according to the sort order of the file. A missing item is imputed the value of the item in the last nonmissing record processed for that imputation class. If the value of an item in the current record is nonmissing it replaces the previous hot-deck value for that imputation class. In this way the hot-deck value for each imputation class is constantly being updated with the value of the last nonmissing case.

3.5.6 Example of SIPP Sequential Hot-Deck Procedure

Figure 3.1 (page 33) illustrates features of the sequential hot-deck procedure used in the SIPP to impute item missing data. The hypothetical data file outlined in Figure 3.1 contains 16 cases which have been sorted by PSU, Segment and Serial Number and allocated into four imputation classes defined by the cross classification of sex (2 levels) and education (2 levels). Two substantive variables (A and B) and an identification variable (Case ID), which represents the concatenation of PSU, Segment and Serial Number, are listed for each case. The substantive variables are dichotomies taking on values of 1 or 0 or "-", which indicates a missing value. The initial starting values in the cold-deck matrix are "1" for variable A and "1" for variable B. For simplicity, the same cold-deck values have been specified for each cell in the classification matrix. In practice, not all cells will receive the same starting values.

The example illustrates several features of the sequential hot-deck procedure: 1) its operational efficiency and simplicity; 2) the initial preprocessing step to update the initial cold-deck values with hot-deck values; 3) when a cold-deck value is used to replace a missing item; 4) the independent imputation of missing items within each imputation class; 5) updating hot-deck values after each nonmissing record is processed; and 6) situations in which a donor is used more than once.

> **Preprocessing to Update the Initial Cold-Deck Values:** The following listing of the 16 hypothetical cases are in sort order and illustrate the preprocessing step which updates cold-deck values with values from the current data set (hot-deck values). Figure 3.1, on the other hand, displays the data in sort order within each classification group to illustrate the subsequent processing step which imputes for item missing data.

The initial cold-deck starting values are "1" for variable A and "1" for variable B. Each record is preprocessed in sort order. The first record encountered with no missing values for both variables A and B updates the cold-deck values. For this example, the second record (Case ID 112) replaces the cold-deck values ("1" and "1") with nonmissing values ("1" and "0"). The nonmissing values supplied by Case ID 112 are now technically known as hot-deck values because they came from the current data set. Each subsequent record with nonmissing data for both variable A and variable B updates the previous hot-deck values, and the first round of updating continues until all records have been processed. For this example, Case ID 413 supplied the final set of updated hot-deck values (0,1). The updated set of hot-deck values are subsequently used as the initial set of hot-deck values in the next step of the imputation procedure which replaces missing values with nonmissing values.

When a Cold-Deck Value is Used to Replace a Missing Item: once the cold-deck values have been updated, the records are allocated into 4 imputation classes based on the values of the classification variables as shown in Figure 3.1, and the portion of the imputation procedure which replaces missing values with nonmissing values begins. In this example no cold-deck values were used to replace (impute) a missing value because the cold-deck values were updated with hot-deck values during the preprocessing step. In practice, few cold-deck values will be used to impute a missing value in the SIPP. To replace a missing value with a cold-deck value in the SIPP requires that all records in a particular section and within the same imputation class have one or more missing items, which is not likely.

Donor Information is Obtained from the Hot-Deck Values: the sequential hot-deck imputation procedure is conducted concurrently, but independently, within each of the imputation classes shown in Figure 3.1. Recall that for this example each imputation class has the same set of initial hot-deck values. For the first imputation class in Figure 3.1 (males with less than a high school education), Variable A for the first case (Case ID 113) is missing and is imputed the initial hot-deck value for variable A ("0"). The initial

Processing the Hypothetical File to Update Cold-Deck Values with Hot-Deck Values from the Current Data Set (Hot-Deck Values)

		Variable A	Variable B
		Initial Cold-Deck Values	
		1	1
Case ID	Imputation Class	Observed Values	
111	3	-	1
112	2	1	0
113	1	-	1
121	3	-	-
122	1	1	0
211	2	0	0
212	1	1	0
213	3	-	-
311	4	0	1
312	3	-	-
313	4	-	-
321	2	0	1
331	2	-	-
411	4	-	0
412	4	-	-
413	1	0	1
		Initial Hot-Deck Values	
		0	1

hot-deck value for Variable B ("1") is updated with the nonmissing value for Case ID 113 ("0"). All other records in the first imputation class have nonmissing values so no further imputations are performed. As each remaining record in the first imputation class is processed, however, the hot-deck values are updated with the values from the current record. The ending hot-deck values for the first imputation class are "0" and "1" for variables A and B, respectively, which were donated by Case ID 413.

Meanwhile, the records in each of the other imputation classes are processed in the same manner: missing values are replaced with hot-

deck values and nonmissing values are updated with current hot-deck values.

Multiple Use of a Donor: the set of cases in the third and fourth imputation classes illustrates instances in which a donor is used more than once due to the sequence of item missing data encountered. For the third imputation class defined by women with less than a high school education, the last three cases, Case IDs 121, 213 and 312, each have missing data for items A and B. As the first case in this imputation class is processed the nonmissing values from Case 111 update the initial hot-deck values for that imputation class (0,1). Note, however, that in this instance the updating simply replaces the initial hot-deck values with the same set of values from Case 111. The hot-deck values supplied by Case 111 are then used to impute the missing items for the remaining cases in this imputation class because no further updating of hot-deck values is possible. In general, a sequential hot-deck procedure will use the same donor *n* times, where *n* is the number of consecutive cases with a missing value on the same item within an imputation class. The possibility that a donor will be used more than once is a disadvantage of the sequential hot-deck procedure, as well as other types of imputation procedures, because when a donor is used more than once the precision of estimates is reduced, in much the same way that weighting reduces the precision of estimates.

3.5.7 Restrictions on Donor Records

Because hot-deck values are replaced item by item, any one record with more than one missing item is not likely to be imputed from the same donor unless the procedure is designed that way. In the SIPP there are sets of closely related items which are jointly imputed from the same donor. (Note, however, that this feature of the sequential hot-deck procedure is not illustrated in Figure 3.1). Generally, these related items on the Core questionnaire are associated with a flashcard shown to respondents in which interviewers are instructed to "mark all that

Figure 3.1 Hypothetical Sample and Illustration of SIPP Sequential Hot-Deck Procedure

		Variable A			Variable B		
Imputation Class	**Case ID**	**Current Replacement Value**	**Observed Value**	**Imputed Value**	**Current Replacement Value**	**Observed Value**	**Imputed Value**
(1) Male: Less than H.S.							
	113	0	-	0	1	0	
	122	0	1		0	1	
	212	1	0		1	0	
	413	0	0		0	1	
(2) Male H.S. and Beyond							
	112	0	1		1	0	
	211	1	0		0	0	
	321	0	0		0	1	
	331	0	-	0	1	-	1
(3) Female: Less than H.S.							
	111	0	0		1	1	
	121	0	-	0	1	-	1
	213	0	-	0	1	-	1
	312	0	-	0	1	-	1
(4) Female: H.S. and Beyond							
	311	0	0		1	1	
	313	0	-	0	1	-	1
	411	0	-	0	1	0	
	412	0	-	0	0	-	0

apply."[13] The advantage of imputing groups of related items from the same donor is that the covariances between the variables are retained.

The unintentional use of a single donor multiple times can present difficulties during the estimation stage because the multiple use of the same donor increases the variance of an estimate. The editing and imputation programs in the SIPP count the number of times the same donor is used. When a single donor is used more than a predetermined number of times, a report is generated and the situation is reviewed and corrective action taken when necessary. The variance of an estimate can also be increased when an extreme value is imputed. In the SIPP, the use of donors with large outlying values is restricted. The restriction is implemented by recoding large outlying values for selected variables prior to imputation. Generally, the presence of large outlying values is confined to asset amounts and not recipiency amounts. The SIPP sequential hot-deck imputation procedure makes no distinction between sample cases whose information was obtained by self or proxy reports and are treated equally as donors.

3.5.8 Imputation Procedures for Control Card and Topical Module Items

Imputation Procedure for Missing Control Card Items: the Control Card is the basic record maintained for each sample household and is the source of several classification variables used to impute item missing data in both the Core questionnaire and Topical Modules. It is essential, therefore, that many Control Card items be nonmissing for all cases. The method for imputing item missing data in the Control Card is also a sequential hot-deck procedure but involves fewer steps than the Core questionnaire item missing data imputation procedure.

The first step in imputing item missing data on the Control Card involves specifying cold-deck values. In the second step the Control Card file is sorted by the same three geographic variables used to sort the Core questionnaire data file: primary sampling area, segment number and serial number. The preprocessing step to identify consistent and nonmissing records and to initially update cold-deck values, and the step which allocates cases into imputation classes in the Core questionnaire imputation procedure is omitted from the Control Card imputation procedure. No imputation classes are maintained in the Control Card procedure because the neighboring household with nonmissing information for an item is considered the best donor available. Another variation from the Core questionnaire procedure is that missing items on the Control Card are replaced with nonmissing values from the same donor, rather than from multiple donors. Once cold-deck values have been specified and the file has been sorted the Control Card records are processed sequentially. Missing items in the first Control Card processed receive cold-deck replacement values. The cold-deck values are subsequently updated with information from the first Control Card record encountered without missing data. Each succeeding Control Card record encountered with no missing information updates the values in the hot-deck matrix. In turn, each Control Card record encountered with missing

[13] Examples of related items in the Core questionnaire which are imputed from the same donor when more than one item is missing include: SC1006-SC1040 (weeks in the reference period looking for work or on layoff) and SC1138-SC1172 (weeks in the reference period absent from work without pay).

information is replaced with nonmissing information from the hot-deck. In this way any missing data on a Control Card record is replaced with information from the nearest neighboring record with no missing data.

Topical Module Imputation Procedure: item missing data in Topical Modules are imputed using the same sequential hot-deck procedure used to impute item missing data in the Core questionnaire. Topical Module data for Type Z and Departure noninterviews are not Type-Z imputed, but rather imputed item by item using the sequential hot-deck procedure used to impute Core questionnaire item missing data. Other features of the implementation of the sequential hot-deck procedure for Topical Modules include: 1) more frequent changes in cold-deck values for variables sensitive to changes in economic activity; and 2) more frequent changes in the composition of classification variables. All other aspects of the Topical Module imputation procedure are similar to the features used to impute item missing data in the Core questionnaire.

3.5.9 Imputation Flags

An imputation flag is associated with each Core questionnaire item subject to statistical imputation. When an item has been imputed using the sequential hot-deck procedure an imputation flag for that item is set to "1"; otherwise, the value of the imputation flag is set to "0." The imputation flags, however, do not indicate whether the imputed values resulted from a refusal or lack of information on the part of the respondent. Imputation flags are not specifically created for missing items which are inferred from the same record, that is items which are logically imputed. If an imputation flag exits for an item and a missing value for that item is inferred from other data on the same record, the imputation flag is not set to "1." Records for Type Z and Departure noninterview cases, which contain imputed data for a majority of items, can be identified by using the indicator variables outlined in Footnote Number 4.

In addition to identifying imputed items, the imputation flags can also be used to calculate the imputation rates for a particular item. The decision to include or exclude imputed values in an analysis or to reimpute a value using another method is the option of the data user. The longitudinal files carry only a subset of imputation flags carried on the wave files, primarily for recipiency and amount variables.[14]

[14] The value of an imputation flag is set during wave processing and is not modified to reflect any changes in a value due to longitudinal editing. See Chapter 4 for a discussion of when a wave imputed value is subsequently modified during longitudinal editing.

3.5.10 Classification Variables, Imputation Matrices and Starting Values for the 1987 SIPP Core Questionnaire Panel

The example of the Core questionnaire sequential hot-deck procedure outlined in Figure 3.1 (Page 33) used the same set of classification variables for each item, but in practice this is not required or even necessarily desirable. Figure 3.2 (Page 38) contains the global set of classification variables used in the 1987 SIPP panel to impute item missing data for selected variables in the Core questionnaire. The classification variables include individual as well as household attributes. Several variables are operationalized into a number of alternative forms. Respondent's age, for example, is recoded in nine alternative forms and marital status has four alternative forms. The set of classification variables come from the current wave Control Card and Core questionnaire.

Tables 3.1 to 3.3 (Pages 44-48) display the arrangement of classification variables into various imputation matrices. Each imputation matrix defines the parameters for imputing item missing data for one or more variables in the Core questionnaire. Table 3.1 contains the imputation matrices for Section 1, "Labor Force and Recipiency"; Table 3.2 contains the imputation matrices for Section 2, "Earnings and Employment"; and Table 3.3 contains the imputation matrices for Section 3, "Amounts and Section 4, "Program Questions." The imputation matrices in Tables 3.1 to 3.3 were used in each wave of the 1987 panel. Table 3.4 (Page 49) contains the set of imputation matrices used for variables measured in Waves 2-8, but not measured in Wave 1. The variables which define each imputation matrix for the 1987 panel are representative of the set of classification variables used in other panels. Variables in the Core questionnaire subject to imputation and the corresponding imputation matrices and cold-deck values for Waves 1-8 of the 1987 panel are contained in Figure 3.3 (Page 50); Figure 3.4 (Page 53) contains comparable information for the 1987 panel for variables measured in Waves 2-8, but not measured in Wave 1.

3.5.11 Rates of Imputation for Selected Core Questionnaire Variables, Waves and Panels

Table 3.5 (Page 54) contains imputation rates for selected variables and waves for the 1988 and 1990 SIPP Panels. The table includes a brief description of the variable, the source code for the variable, which can be used to determine the full text and context of the variable in the SIPP questionnaire or to obtain information about the variable in SIPP technical documentation, and the base upon which the imputation rate is calculated. The variables in the tables were selected to represent different content area in the Core questionnaire as well as to indicate the range of imputation rates. An examination of the imputation rates in Table 3.5 shows that the largest imputation rates are associated with amounts, and that, in general, rates of imputation for a particular item are stable across waves and between panels.

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire

AGE

	AGE1	AGE2	AGE3	AGE4	AGE5	AGE6	AGE7
1.	15 to 19	Under 17	Under 25	Under 25	Under 25	Under 61	Under 35
2.	20 to 24	17 to 22	25 to 34	25 to 44	25 to 44	61 to 64	35 to 54
3.	25 to 39	23 to 24	35 to 44	45 to 61	45 to 64	65 to 69	55 and over
4.	40 to 54	25 to 29	45 to 54	62 to 64	65 and over	70 and over	
5.	55 to 64	30 to 49	55 to 64	65 to 69			
6.	65 and over	50 and over	65 and over	70 and over			

	AGE8	AGE9
1.	Under 25	17 to 22
2.	25 to 44	23 to 24
3.	45 and over	25 to 29
4.		30 to 49

RACE

1. White and other nonblack
2. Black

SEX

1. Male
2. Female

EDUCATIONAL ATTAINMENT

1. Less than 12 years
2. 12 to 15 years
3. 16 years or more

MARITAL STATUS

	MS1	MS2	MS3	MS4
1.	Married	Married	Married	Ever widowed or currently widowed
2.	Widowed	Widowed	Separated or divorced	All others
3.	Divorced	Separated or divorced	Widowed or never married	
4.	Separated	Never married		
5.	Never married			

DISABILITY STATUS

	DS1	DS2
1.	Has a disability	Has a service connected disability
2.	Other	No service connected disability; DK

WORK EXPERIENCE

	WE1	WE2	WE3
1.	Worked less than 4 weeks	Did not work: going to school	Worked 1 or more weeks
2.	Worked 5 to 8 weeks	Did not work: other situations	Did not work
3.	Worked 9 to 12 weeks	Worked 1 or more weeks	
4.	Worked 13 weeks or more		

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire (Continued)

MONTHS WITH JOB

	MJ1	MJ2		MJ3
1.	Entire period	None		(1-15) Each unique combination of months with a job
2.	Part of the period	One	(16)	Did not work
3.	Did not work	Two		
4.		Three		
5.		Four		

WORKER STATUS

	WS1	WS2
1.	Full-time worker	Worker
2.	Part-time worker	Nonworker
3.	Nonworker	

USUAL HOURS WORKED

1. Under 20 hours
2. 20 to 34 hours
3. 35 hours or more

WORK EXPERIENCE OF RECIPIENT OF INCOME TYPE

	WER1	WER2
1.	Did not work during reference period	Did not work during reference period
2.	Worked full time	Worked one or more weeks
3.	Worker part time	

WEEKS LOOKING FOR WORK OR ON LAYOFF

1. None
2. 1 to 4 weeks
3. 5 to 9 weeks
4. 10 or more weeks

NUMBER OF EMPLOYEES

1. 1 or 2
2. 3 to 14
3. 15 to 49
4. 50 or more
5. NIU

REGULAR SALARY RECIPIENCY

1. Yes
2. No

TYPE OF BUSINESS

1. Incorporated
2. Sole proprietorship
3. Partnership
4. NIU

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire (Continued)

RETIREMENT STATUS

1. Both spouses retired
2. This spouse only retired
3. Neither retired or this person not retired

OWN CHILDREN PRESENT

	CHILD1	CHILD2	CHILD3
1.	No children under 18	No children under 18	No children
2.	One or more children under 18	One child under 18	No children under 5
3.		Two or more children under 18	One or more children under

	CHILD4
1.	One or more children under 16
2.	All children are 16 to 17

SIZE OF HOUSEHOLD

1. One
2. Two
3. Three
4. Four
5. Five or more

TYPE OF PUBLIC HOUSING

1. Owed by local housing authority
2. Lower rent pubic housing

HOUSEHOLD RELATIONSHIP

	HHREL1	HHREL2
1.	Householder	Family householder
2.	Spouse of householder	Nonfamily householder or secondary unrelated individual
3.	Other household member	Spouse of householder
4.		Other related household member

RELATIONSHIP

1. Reference person, living with relatives
2. Reference person, living alone or with only nonrelatives
3. Spouse of reference person
4. Child of reference person
5. Other relative of reference person
6. Nonrelative of reference person, but related to others in household
7. Nonrelative of reference person, and not related to anyone else in household

VETERAN STATUS

1. Vietnam era veteran
2. May 1975 or later
3. Other eras

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire (Continued)

SOCIAL SECURITY RECIPIENCY

1. Received social security
2. Did not receive social security

INCOME RECIPIENCY (SS, SSI, VA)

1. Received social security, supplemental social security or veterans's administration benefits
2. Did not receive income noted in "1"

RECIPIENCY TYPE

1. Joint (social security or RR only)
2. Not joint

SPOUSE RECIPIENCY: ANY OF ISS CODES 100-110, 120, 130, 140, 150, 174

1. Yes
2. No
3. Not answered
4. No spouse present

RECEIVED ANY OF ISS CODES 100-110, 120, 130, 140, 150, 174

1. Yes
2. No

OTHER INCOME RECIPIENCY: ANY OF ISS CODES 2, 9, 10, 13, 30-36, 38

1. Yes
2. No

RECEIPT OF ASSISTANCE (4-MONTH RECIPIENCY)

1. Received on one more of the following:
   AFDC
   Food Stamps
   Medicaid coverage
   WIC
2. Did not receive one or more of these types

ADULT MEDICAID COVERAGE

1. Person covered
2. Person not covered

NUMBER OF PERSONS COVERED BY INCOME TYPE

	COVERED1	COVERED2	COVERED3
1.	NIU	NIU	One
2.	One	One	Two
3.	Two	Two	Three or more
4.	Three	Three	
5.	Four	Four or more	
6.	Five		
7.	Six or more		

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire (Continued)

PERSON'S EARNINGS IN FOUR-MONTH PERIOD

	EARN1	EARN2
1.	No earnings	None
2.	Under $3,000	Under $1,500
3.	$3,000 to $6,999	$1,500 to $2,999
4.	$7,000 to $9,999	$3,000 to $4,999
5.	$10,000 to $17,499	$5,000 to $9,99
6.	$17,500 and over	$10,000 and over

HOUSEHOLD INCOME (EXCLUDING PROPERTY INCOME)

1. Under $1,500
2. $1,500 to $2,999
3. $3,000 to $5,999
4. $6,000 or more

HOUSEHOLD INCOME SOURCES

1. Households with one or more of the following income sources:
   Federal SSI
   AFDC
   Food Stamps
   WIC
   Medicaid
2. Households without one or more of the specified income sources

INDUSTRY

1. Agriculture, forestry, and fisheries (Codes 010-031)
2. Mining (Codes 040-050)
3. Construction (Code 060)
4. Manufacturing (Codes 100-392)
5. Transportation, communications, and other public utilities (Codes 400-472)
6. Wholesale trade (Codes 500-571)
7. Retail trade (Codes 580-691)
8. Finance, insurance, and real estate (Codes 700-712)
9. Business and repair services (Codes 721-760)
10. Personal services (Codes 761-791)
11. Entertainment and recreation services (Codes 800-802)
12. Professional and related services (Codes 812-892)
13. Public administration (Codes (900-932)
14. Armed Forces (Code 991)

OCCUPATION

1. Executive, administrative, and managerial (Codes 003-037)
2. Architects, engineers, mathematical and computing scientists, natural scientists, social scientists, and urban planners (Codes 043-083,166-173)
3. Health diagnosing occupations, lawyers, and judges (Codes 084-089, 178-179)
4. Teachers, counselors, librarians, archivists, curators (Codes 113-165)
5. Other professional specialty occupations (Codes 095-106, 174-177)
6. Health technicians (Codes 203-208)
7. Engineering and science technicians (Codes 213-225)
8. Other technicians (Codes 226-235)

Figure 3.2 Classification Variables for Imputing Item Missing Data in the 1987 SIPP Panel Core Questionnaire (Continued)

OCCUPATION (Continued)

9. Sales supervisors, proprietors, engineers, and representatives (Codes 243-259)
10. Cashiers (code 276)
11. Other sales occupations (263-275, 277-285)
12. Administrative support supervisors (Codes 303-307)
13. Secretaries and stenographers (Codes 313-315)
14. Other administrative support occupations (Codes 308-309, 316-389)
15. Protective service occupations (Codes 413-427)
16. Health and personal service occupations (Codes 445-4447, 456-469)
17. Private household, cleaning, and building service occupations (Codes 403-407, 448-455)
18. Food preparation and service occupations (Codes 433-444)
19. Mechanics and repairers (Codes 503-549)
20. Construction trades and extractive occupations (Codes 553-617)
21. Precision production occupations (Codes 633-699)
22. Machine operators, assemblers, and inspectors (Codes 703-799)
23. Transportation and material moving occupations (Codes 803-859)
24. Supervisors, handlers, equipment cleaners, and laborers (Codes 863, 869, 875-889)
25. Helpers (Codes 864-867, 873)
26. Farm operators and managers (Codes 473-476)
27. Other agricultural and related occupations, forestry and logging, fishers, hunters, and trappers (Codes 477-499)
28. Armed Forces (Code 905)

OCCUPATION/INDUSTRY

1. Farmers (Codes 473, 474)
2. Health diagnosing occupations, lawyers (Codes 084-089, 178)
3. Agricultural industries (Codes 010-021)
4. Construction industries (Code 060)
5. Manufacturing industries (Codes 100-392)
6. Communications (Codes 400-472)
7. Wholesale trade (Codes 500-571)
8. Retail trade (Codes 580-691)
9. Finance, insurance, and real estate (Codes 700-712)
10. Mining, forestry, and fisheries (Codes 030-050)
11. Business services (Codes 721-750)
12. Repair services (Codes 751-760)
13. Personal, entertainment and recreation services and private household workers (Codes 761-802)
14. Professional and related services (Codes 812-892)

    Dummy cells:
15. Public administration (Codes (900-932)
16. Armed Forces (Code 991)

Table 3.1 Imputation Matrices for Section 1: Labor Force and Recipiency (Waves 1-8, 1987 Panel)

Classification Variables	Matrices																										
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Age																											
AGE1			■																						■		
AGE2																										■	
AGE3	■	■		■		■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■			■
AGE6					■																						
Race	■	■	■		■	■	■	■	■	■	■	■	■	■	■	■	■								■	■	■
Sex	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■	■
Marital Status																											
MS1																		■	■					■			
MS2					■									■													
MS3	■	■					■	■			■		■		■	■	■										
Disability Status																											
DS1	■	■			■								■														
DS2				■																							
Own Children Present																											
CHILD1	■	■																■						■			
CHILD2									■		■			■	■	■	■										
Work Experience																											
WE1		■																									
WE2																										■	
WE3			■						■		■			■													

Table 3.1 Imputation Matrices for Section 1: Labor Force and Recipiency (Waves 1-8, 1987 Panel, Continued)

Classification Variables	Matrices																									
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Worker Status																											
WS1																											
WS2																											
Weeks Looking or on Layoff																											
Household Relationship																											
HHREL1																											
Veteran Status																											
Social Security Recipiency																											
Income Recipiency (SS,SSI,VA)																											
Adult Medicaid Coverage																											
Relationship																											
Months with Job																											
MJ1																											
MJ2																											
MJ3																											
Spouse Recipiency (ISS codes 100-110, 120, 130, 140, 150, 174)																											
Receipt of Assistance																											

Table 3.2 Imputation Matrices for Section 2: Earnings and Employment (Waves 1-8, 1987 Panel)

Classification Variables	Matrices								
	28	29	30	31	32	33	34	35	36
Age									
AGE8	▮	▮	▮	▮					
Race	▮	▮	▮						
Sex	▮	▮	▮	▮	▮	▮	▮	▮	▮
Educational Attainment	▮	▮	▮	▮	▮	▮		▮	
Industry		▮							
Occupation			▮						
Industry/Occupation				▮	▮	▮	▮	▮	▮
Usual Hours Worked				▮		▮		▮	▮
Number of Employees					▮			▮	
Type of Business						▮	▮		
Regular Salary Recipiency							▮		

Table 3.3 Imputation Matrices for Section 3: Amounts and Section 4: Program Questions (Waves 1-8, 1987 Panel)

Classification Variables	Section 3										Section 4		
	37	38	39	40	41	42	43	44	45	46	47	48	49
Age													
AGE1													
AGE4													
AGE5													
AGE7													
Race													
Sex													
Educational Attainment													
Marital Status													
MS4													
Disability Status													
DS1													
Presence of Own Children													
CHILD3													
CHILD4													
Recipiency Type													
Retirement Status													
Size of Household													
Work Experience													
WER1													
WER2													
Worker Status													
WS1													
WS2													
Household Relationship													
HHREL1													
HHREL2													

Table 3.3 Imputation Matrices for Section 3: Amounts and Section 4: Program Questions (Waves 1-8, 1987 Panel, Continued)

Classification Variables	Section 3										Section 4		
	37	38	39	40	41	42	43	44	45	46	47	48	49
Weeks Looking or on Layoff													
Person's 4-Month Earnings													
EARN1													
EARN2													
Number of Persons Covered													
COVERED1													
COVERED2													
COVERED3													
Household Income													
Type of Public Housing													
Household Income Source													

Table 3.4 Imputation Matrices for Section 1: Labor Force and Recipiency (Waves 2-8, 1987 Panel)

Classification Variables	Matrices							
	50	51	52	53	54	55	56	57
Age								
Age1								■
Age3	■		■			■	■	
Age6				■	■			
Age9		■						
Race	■	■	■	■	■	■	■	■
Sex	■	■	■	■	■	■	■	■
Marital Status								
MS2	■		■	■	■	■	■	
Disability Status								
DS1	■							
Work Experience								
WE2		■						
WE3	■		■					
Worker Status								
WS1		■						■
Other Income Recipiency (ISS codes 2, 9, 10, 13, 30-36, 38)						■		
Income Recipiency (SS,SSI,VA)				■				
Spouse Received Any of ISS Codes: 100-110, 120, 130, 140, 150, 174								■
Received Any of ISS Codes: 100-110, 120, 130, 140, 150, 174								■

Figure 3.3 Core Questionnaire Items Subject to Imputation and Corresponding Imputation Matrices (Waves 1-8, 1987 Panel)*

Table 3.1: Section 1, Labor Force and Recipiency

Variable	Matrix	Cold-Deck Values	Variable	Matrix	Cold-Deck Values	Variable	Matrix	Cold-Deck Values
SC1002	1	2	SC1416	9	2	SC1636-1646	25	0, SC1636=1
SC1004-1040	1	0, SC1004=1				SC1648	25	2
SC1042	1	2	SC1418	10	2	SC1650	25	2
SC1044	1	4				SC1652	25	2
SC1046	1	2	SC1422	11	2	SC1654	25	2
SC1048-1054	1	1						
SC1056	1	1	SC1426	12	2	SC1656	26	3
SC1058	1	2	SC1428-1452	12	0, SC1434=1	SC1658-1666	26	0, SC1658=1
SC1060-1096	1	0, SC1060=1	SC1456	12	3	SC1668	26	3
SC1098	1	7				SC1670	26	2
SC1100-1134	1	0, SC1100, SC1102=1	SC1462	13	2	SC1672-1692	26	0, SC1690=1
SC1136	1	2	SC1472	13	1			
SC1138-1172	1	SC1138-1140=1 SC1142-1172=0				SC1696	27	2
SC1174	1	7	SC1480	14	2			
SC1176	1	2	SC1484	14	2			
SC1178-1214	1	0, SC1180-1182=1	SC1486-1498	14	0, SC14888=1			
SC1216	1	2						
SC1218	1	4	SC1502	15	2			
SC1220	1	2						
			SC1508	16	2			
SC1222-1228	2	0, SC1222=1						
SC1230	2	40	SC1526	17	2			
SC1232	2	2	SC1528-1534	17	0, SC1528=1			
SC1233-1237	2	0, SC1233=1						
SC1238	2	6	SC1536	18	2			
SC1240	3	2	SC1537	19	2			
SC1246	3	2						
			SC1538	20	1			
SC1338	4	2						
			SC1540-1546	21	0, SC1540=1			
SC1342	5	2						
SC1352	5	2	SC1548	22	1			
SC1354	5	2						
			SC1550	23	1			
SC1360	6	2						
SC1362	6	2	SC1552	24	1			
SC1364-1380	6	0, SC1366=1						
SC1382	6	2	SC1624	25	1			
			SC1626	25	1			
SC1386	7	2	SC1628	25	2			
			SC1630	25	2			
SC1388	8	2	SC1632	25	2			
SC1390-1412	8	0, SC1398=1	SC1634	25	2			

* Core questionnaire variables in Figure 3.3 are identified by their source code (SC) numbers.

Figure 3.3 Core Questionnaire Items Subject to Imputation and Corresponding Imputation Matrices (Waves 1-8, 1987 Panel, Continued)

Table 3.2: Section 2 Earnings and Employment

Variable	Matrix	Cold-Deck Values
Occupation	28	019, 019*
Industry	28	641, 601*
SC2012	29	1
SC2024	30	6
SC2026	30	2
SC2028	30	800
SC2030	30	1
SC2032-2038	30	800
SC2044	30	1
SC2046	30	1
SC2214	31	1
SC2218	31	10
SC2220	32	2
SC2222	32	2
SC2232	33	1
SC2234	34	2
SC2238-2244	35	350
SC2254	35	1
SC2256	35	6300 or -3150
SC2260	36	500

Table 3.3: Section 3 Amounts

Variable		Matrix	Cold-Deck Values
SC3008		37	1
SC3018-3030			
for: SC3000=	1	38	400
	2	38	550
	3	38	250
	4	38	50
	9	38	350
	30	38	300
	31	38	800
	32	38	800
	33	38	300
	34	38	400
	35	38	400
	36	38	100
	37	38	500
	38	38	200
SC3018-3030			
for: SC3000=	50	39	100
	51	39	200
	52	39	500
	53	39	150
	54	39	150
	55	39	100
	56	39	200
SC3018-3030			
for: SC3000=	20	40	300
	21	40	150
	22	40	200
	23	40	250
	24	40	100
	28	40	200
	29	40	400

Table 3.3: Section 3 Amounts (Continued)

Variable		Matrix	Cold-Deck Values
SC3018-3030			
for: SC3000=	5	41	400
	6	41	200
	7	41	200
	10	41	500
	11	41	350
	12	41	400
	13	41	500
SC3018-3030			
for: SC3000=	8	42	200
	40	42	300
	41	42	300
SC3072-3084			
for: SC3000=	1**	43	200
	2**	43	200
SC3124-3136			
for: SC3000=	27	40	100
SC3138-3144			
for: SC3000=	25	44	1

* The first cold-deck value listed is for wage and salary earners; the second cold-deck value is for self-employed persons.

** Child's payments only.

Figure 3.3 Core Questionnaire Items Subject to Imputation and Corresponding Imputation Matrices (Waves 1-8, 1987 Panel, Continued)

Table 3.3: Section 3 Amounts (Continued)

Variable	Matrix	Cold-Deck Values
SC4310	45	1
SC4312	45	50
SC4410	45	1
SC4412	45	100
SC4504	45	100
SC4516	45	100
SC4602	45	1
SC4604	45	1200
SC4606-4608	45	300 OR -150
SC4710	45	1
SC4712	45	1200
SC4318	46	1
SC4320	46	25
SC4418	46	1
SC4420	46	50
SC4500	46	1
SC4508	46	50
SC4512	46	2
SC4518	46	50
SC4610	46	2
SC4612	46	800
SC4614-4616	46	200 OR -100
SC4618	46	2
SC4620-4622	46	600 OR -300
SC4714	46	2
SC4716	46	800
SC4720-4722	46	400 OR -200

Table 3.3: Section 4 Program Questions

Variable	Matrix	Cold-Deck Values
SC4804 OR SC4810	47	200
SC4806 OR SC4812	47	2
SC4814	48	300
SC4816	49	2
SC4818-4822	49	0, SC4822=1
SC4824	49	200
SC4828	49	1
SC4830	49	5
SC4832	49	2
SC4834	49	5
SC4836-4838	49	SC4836=0; SC4838=1
SC4840	49	2
SC4842	49	5
SC4844-4846	49	SC4844=0; SC4846=1

Figure 3.4 Core Questionnaire Items Subject to Imputation and Corresponding Imputation Matrices (Waves 2-8, 1987 Panel)

Cover Page

Variable	Matrix	Cold-Deck Values
SC902	7	2

Table 3.4: Section 1
Labor Force and Recipiency

Variable			Matrix	Cold-Deck Values
SC1254-1282	for: SC1252-SC1280=	5-7,10	3	2
		1,3,4	5	1
		8,9,11,12	50	1
		13	50	2
		2,30-35	8	1
		36-38	8	2
		20,23,25,27-29	14	1
		21,22,24,50,51	14	2
		40	51	2
		52-56	52	2
SC1284			53	2
SC1286-1294			54	0, SC1286=1
SC1296			55	2
SC1298-1322			56	0, SC1306=1
SC1502			15	2
SC1504			15	2
SC1592-1620	for: SC1590-1618=	100-107,110,	25	1
		120,130,174	25	1
		140,150	25	2
SC1622			57	2
SC1626-1654			25	0, SC1626=1

Table 3.5 Item Missing Data Imputation Rates, 1988 and 1990 SIPP Panels: Selected Variables and Waves

Variable	Source Code	1988 Panel				1990 Panel			
		Wave 2		Wave 6		Wave 1		Wave 3	
		%	Base	%	Base	%	Base	%	Base
Household Level:									
Monthly Rent	SC4804	28.8	580	30.5	521	27.5	1,206	31.5	1,155
Energy Assistance	SC4816	3.6	11,737	3.9	11,549	2.2	21,898	3.5	21,950
School Lunch	SC4828	6.5	3,909	8.8	3,677	5.8	7,093	7.9	7,147
Apply for School Lunch	SC4834	6.9	2,276	9.3	2,202	5.0	4,443	7.3	4,377
Person Level:									
Wage and Salary Employment									
Occupation/Industry	SE-OCC/IND	0.6	15,031	1.2	15,145	0.9	27,539	1.7	28.284
Last Month's Wage	SC2032	8.6	14,972	3.5	15,059	3.8	27,446	3.8	28,178
Self Employment									
Occupation/Industry	SE-OCC/IND	1.4	2,065	2.6	2,065	0.5	3,706	4.2	3,561
Business Income Last Month	SC2238	17.5	1,539	21.4	1,554	16.3	2,778	20.2	2,696
Work: Looking/on Layoff	SC1002	0.6	8,114	0.7	8,096	0.1	16,045	0.8	15,338
Weeks Looking/on Layoff	SC1004-1040	8.4	633	13.3	482	8.8	1,333	10.7	1,210
Hours Work on Job	SC1230	1.3	15,245	1.3	15,400	1.2	28,490	1.3	28,655
Covered by Medicaid?	SC1502	1.1	21,047	1.1	21,252	0.4	42,063	1.1	39,469
Covered by Health Insurance	SC1536	0.2	23,400	0.2	23,496	0.2	44,535	0.2	43,993
Enrolled in School	SC1656	0.1	23,400	0.1	23,496	0.1	44,535	0.1	43,993
Social Security Amount	SC3018	13.4	9,634	15.5	9,855	11.7	18,346	15.3	18,605
Food Stamp Amount	SC3124	6.0	739	8.5	703	4.4	1,614	6.4	1,685
Type of Veterans' Benefits	SC3058	6.3	397	6.9	375	11.0	628	10.5	630
Asset Codes 100, 101, 102, 103: Savings, Money Market, CD's, Interest Earning Checking									
Own Jointly with Spouse	SC4310	0.6	5,103	1.0	5,100	1.7	9,170	0.8	8,835
Earned Interest: Last 4 Months	SC4312	34.2	4,425	36.8	4,402	34.1	7,916	38.6	7,613
Asset Codes 104, 105, 106, 107: Money Market Funds, US Government Securities, Municipal or Corporate Bonds, Other									
Own Jointly with Spouse	SC4410	0.3	639	0.5	665	2.0	1,173	0.8	1,119
Earned Interest: Last 4 Months	SC4412	47.4	439	55.3	432	48.7	794	53.8	768
Asset Code 110: Stocks or Mutual Fund Shares									
Joint Dividend Check Amount	SC4504	20.0	979	21.6	940	22.0	1,770	24.5	1,652

4. Longitudinal Edits

4.1 Introduction

The editing and imputation procedures described in previous chapters are independently applied to SIPP data within a given wave in order to expedite the availability of microdata files for public use. Users who want to analyze SIPP data longitudinally can link records across waves but the procedure can be time consuming and expensive. To facilitate analysis of SIPP data across waves, the Bureau of the Census has developed a system which links together wave records to produce longitudinally processed data sets. The longitudinal edits are applied only for selected variables and only after all waves of a panel have been processed cross sectionally. This chapter provides an overview of the procedures which edit the data for consistency over time to produce the SIPP Full Panel Microdata Research files. Throughout the chapter the Full Panel Microdata Research file is referred to as the longitudinal file.

Technically, the longitudinal consistency edits reviewed in this chapter are different from the statistical matching and imputations described in previous chapters. The longitudinal consistency edits do not replace missing data in one case with reported data from another case. Rather, when a data value is modified during longitudinal editing the replacement value is obtained: 1) from the same or different wave for the same case; 2) by extrapolation from a previous wave or by interpolation between waves for the same case; or 3) by some other procedure such as averaging which evens out fluctuations in a series of imputed values. In order to make the longitudinally processed file consistent across waves and to take advantage of information reported in other waves, these procedures can lead to modifications in both reported and imputed values for one or more waves. In this chapter the term "longitudinal edits" is used to refer to any modifications made to data during longitudinal processing rather than using the term longitudinal imputation. When a data value is modified during longitudinal editing the value of an existing imputation flag is not changed.[15] Also, Changes made during longitudinal editing are not reflected in the cross-sectional wave files.

The longitudinal data sets are constructed in several steps, each of which is performed independently on a subset of variables. The four subsets of variables which are independently processed consist of:

* Demographic and household composition variables;

* Wage and salary and self-employment variables;

[15] Changes in reported data values resulting from longitudinal consistency edits are not flagged in the longitudinal record. If a value imputed during cross-sectional processing is subsequently changed during longitudinal editing, the value of the imputation flag, if present, is not altered from the value carried from individual wave files. Very few reported values are changed, however.

* Income sources 1-56 and 100-150; AFDC, Food Stamps and WIC variables (to eliminate double counting); program coverage variables; and

* Health and medical coverage variables

Each subset of variables is processed in a three-step sequence. First, the relevant variables for each section are extracted from the individual wave files and then moved to a record constructed for each sample person. Second, the longitudinal edits are applied. Third, the edited data are added to the longitudinal file which is constructed in segments by joining together each subset of edited variables. The longitudinal edits associated with each subset of variables are discussed in the following sections.

4.2 Goals of Longitudinal Edits

The longitudinal editing procedures are guided by several considerations, including:

* The fundamental requirement to ensure cross-wave consistency, which only becomes apparent when multiple waves of SIPP data are examined together;

* The realization that not all possible edits and consistency checks can be implemented;

* The opportunity to address any problems associated with wave files;

* The preference to replace imputed values from one wave with reported values from another wave when available; and

* The need to reduce the number of variables carried from the wave files to the longitudinal files in order to condense the physical size of the data sets. For example, variables which should be unchanging, such as sex and race, are carried once. Other variables carried as a series of related items on the wave files, such as health insurance, are carried as summaries on the longitudinal files.

4.3 Longitudinal Edits for Demographic and Household Composition Variables

A limited amount of longitudinal editing is accomplished for some demographic variables during wave processing. For other demographic variables, and all household variables, inconsistencies are detected and corrected when the data are processed longitudinally. A Control file is developed at Wave 1 which contains a unique identifier for each sample person as well as the individual's age, sex and race. This file is used in subsequent waves to control the receipt of data from the field. Although the control receipt system was developed primarily to guarantee the validity of person identifiers across interviews, it also provides a limited means of detecting inconsistencies in age, sex and race across waves.[16] As each wave of data is received, the reported age, sex and race of the sample person is checked against the control receipt file and any corrections made. This system cannot detect all inconsistencies in age, sex and race, however. Errors made in data collection in Wave 1, the wave from which the Control file is derived, for example, are corrected on the Control file as they are discovered but usually too late for the same corrections to be applied to the cross-sectional file before they are released for public use. These remaining inconsistencies are handled as part of the longitudinal editing process.

Other demographic variables have no longitudinal editing component in the wave processing and inconsistencies in these variables are also addressed in the longitudinal edits. For example, persons reported as widowed in Wave 1 may be reported as never married in Wave 2, or two persons reported as parent and child in one wave may be reported as husband and wife in another wave. These and many other inconsistencies only become apparent when multiple waves of data are examined together.

Household composition variables are also edited for consistency during longitudinal processing. For example, a household may be classified as group quarters in one wave and as a housing unit in another wave. Variations in the classification of housing units across waves, even though address and household composition are unchanged, affect the treatment of respondents. The usual "cross sectional" Census definitions used in editing basic demographic characteristics require that group quarters be occupied by unrelated individuals. During the months that respondents occupy a group quarters they are forced by the cross-sectional edit to be unrelated individuals; for the other months they may be shown as parent-child, husband-wife etc. To correct this inconsistency, a longitudinal edit requires later waves to be consistent with the data reported in Wave 1. Wave 1 is used as the standard against which inconsistencies are judged because the panel weight is based on characteristics as reported in Wave 1. Consequently, when a choice has to be made about which wave has correct data it is preferable to avoid changes in Wave 1 characteristics.

The demographic characteristics affected by the longitudinal relationship/composition consistency edits are the following:

---

[16] Because the person identifier variables are subject to strict controls during the data collection phase no cross sectional or longitudinal editing of this variable is required.

* Relationship to household reference person;
* Age;
* Race;
* Sex;
* Marital status;
* Family type;
* Relationship to family reference person;
* Family number;
* Person number of parent;
* Person number of spouse;
* Reasons for entry into or exit from a housing unit;
* Dates of entry into or exit from a housing unit;
* Identifiers of households to which each person belongs; and
* Type of living quarters.

4.4 Longitudinal Edits for Labor Force Variables

Longitudinal Consistency Edits for Labor Force Activity Variables: the Core questionnaire for each wave begins with a series of items covering various aspects of each sample person's labor force situation during the four-month reference period (items 1 through 8d; SC1000-SC1238), and is asked independently in each wave; that is, information from previous interviews is not referenced in the current interview. This series of questions is not longitudinally edited, although a cross-sectional edit was introduced in the process of creating the 1984 panel file. These labor force participation questions are not longitudinally edited because: 1) a nonmissing response is required in the first item (whether or not the sample person worked during the four-month reference period: SC1000) in order for the interview to be considered complete; and 2) item missing data rates for other key status indicators are low (generally less than 1 percent). Longitudinal edits for the number of weeks in each month with a particular status, such as without pay, looking for work or on layoff, may be implemented in the future to improve the chronology of these occurrences across waves. The item missing data rates for these items average around 10 percent.

The cross-sectional edit introduced as a result of developing the longitudinal file examines the consistency between weeks with a job or business recorded in the Labor Force Activity section and weeks employed by specific employers reported in the Earnings and Employment section of the questionnaire. The edit is achieved in three steps. The first step determines the total number of weeks a sample person was employed with all employers for each month using data from the Earnings and Employment section. The second step compares the value of these weeks with the value reported in the Labor Force Activity section. When the two values do not agree, the value derived in the Labor Force Activity section is edited to agree with the value derived from the Earnings and Employment section. The third step adjusts corresponding labor force activity items such as weeks without pay and weeks looking for work or on layoff to be consistent with the field containing the edited number of weeks with a job or business. Table 4.1 displays the percentage distribution of cases in which the number of weeks with a job or business was in agreement between the two sections, the

percentage of cases which required editing and the nature of the edit for the 1984 longitudinal panel file. The longitudinal file contains separate values for each of the following labor force variables:

* Employment status recode;
* Number of weeks with a job or business;
* Number of weeks without pay; and
* Number of weeks looking for work or on layoff;

Table 4.1 Percentage Distribution of Selected Aspects of the Number of Weeks with a Job Edit: 1984 SIPP Longitudinal File

Situation	Percentage Distribution of Occurrences
Total months checked for correspondence	100.0%
No edits required	96.9
Number of weeks in two sections inconsistent and edit required	3.1
Weeks with a job or business changed to 0	.5
Weeks with a job or business changed to 1-5	2.6

Longitudinal Consistency Edits for Job or Business ID Number: the SIPP Core questionnaire collects data on up to two wage and salary jobs and two self-employment businesses. The system which identifies different jobs held or self-employment businesses owned by a sample person during a panel is based on assigning a number from 1 to N to each job or business as it first appears during the panel. These ID numbers are subsequently used to link data about a particular job or business within and between waves. Errors in assigning ID numbers typically occur when a sample person changes jobs. The ID number may be correctly assigned to the new job in the current reference period, but incorrectly assigned in the subsequent reference period.[17] More complex problems can occur when the number of employers or businesses reaches three or more.

[17] For example, a sample person starts a new job sometime after the start of the reference period. The old job is assigned "1" and the new job is assigned "2" for the current reference period. In the subsequent reference period, the previously new job is incorrectly assigned a "1" instead of the correct ID number, "2."

The purpose of this edit is twofold. First, the edit corrects obvious inconsistencies in the assignment of job or business ID numbers to prevent linking together data about different jobs or businesses. Second, the edit identifies jobs or business with imputed earnings amounts and replaces the imputed values with reported amounts obtained in previous or subsequent interviews. Table 4.2 shows the results of the edit of job identification number for the 1984 longitudinal file.

Table 4.2 Number of Consistency Edits for Job or Business Identification Number: 1984 SIPP Longitudinal File

Situation	Occurrences n	%
Total number of employer records	178,805	100.0
Records requiring edit of job ID	7,661	4.3
Job ID of first employer record not "1"	626	0.4
Gaps in job ID's	3,629	2.0
Job ID assigned incorrectly	3,406	1.9

Longitudinal Edits for Hourly Wage and Monthly Earnings Amounts: the edits for hourly wage and monthly earnings amounts are performed after the job or business ID numbers have been edited. Imputed hourly wage rates are replaced with the average of the reported values for a specific employer if at least one reported value is present. If no reported values are available for a specific job, the imputed values are replaced by the average imputed value. When an imputed hourly wage rate for a specific job is replaced with the average of the reported or imputed values, monthly amounts earned at that job must also be recalculated. The monthly amount earned for hourly wage jobs is calculated by multiplying the number of weeks with pay for that month by: 1) the usual number of hours worked per week; and 2) the edited hourly wage rate for that month.

The edit procedure for earnings amounts collected on a monthly basis is also based on an averaging algorithm which results in replacement of imputed monthly earnings values with either values derived from reported data, or with values derived from all cross-sectionally imputed values, if no reported data exist. The first step in the edit procedure involves calculating an "implied" hourly wage and salary amount for a specific job. The implied hourly wage amount is calculated by first replacing imputed monthly earnings amounts with either the average of the reported amounts, or if no reported amounts are present, by the average of the imputed amounts. Months with zero earnings are excluded from the calculation. The monthly earnings amounts are then summed and divided by the sum of the products of: 1) the number of weeks with pay; and 2) the usual hours worked per week for each month. The quotient is the implied hourly wage and salary amount. The replacement value for imputed monthly earnings amounts is obtained by multiplying the implied hourly wage rate by: 1) the number of weeks with pay; and 2) the usual number of hours worked per week for the month.[18]

[18] Note that the usual hours worked per week is reported once for the four-month reference period; therefore, the same figure is used for each month of a specific reference period for a specific employer.

An additional edit is performed on earnings amounts collected on a monthly basis for workers paid by the hour. This edit compares the reported monthly earnings amount with a calculated monthly earnings amount.[19] If the reported monthly amount is 10 times greater than the calculated amount, the reported amount is replaced with the calculated amount. The purpose of this edit is to decrease the number of monthly amounts that have a high probability of being wrong.

4.5 Longitudinal Edits for Income Sources 1-56 and 100-150

The longitudinal edits for general amount variables are described separately for: 1) nonwage and salary income sources numbered 1-56; and 2) asset types numbered 100-150.[20] Also described in this section are edits applied to program coverage variables and edits designed to detect the presence of duplicate amounts.

Income sources and amounts numbered 1-56 are not directly acquired from labor market activity and include state and local transfer programs, public and private pension and retirement programs, annuities, trusts and so on. Asset types, and income from asset types numbered 100-150, include real property, royalties and financial instruments such as checking accounts, stocks and bonds and so on.

Edits for Income Sources 1-56: the income profile of each household member age 15 or over is established in the initial Wave 1 interview and updated during each subsequent interview. Unlike the employer and earnings data, the collection of income data for sources numbered 1-56 is not independent from one interview to the next. Questions regarding the receipt of specific sources of income for a current reference period are preceded by the interviewer reading a list of income sources reported as received in the previous reference period. The review of income sources received in the previous reference period provides the opportunity to identify errors and to update the list of income sources received for the current reference period.[21] The collection of amount data received from various sources, however, is independent from one interview to the next. Amount data is collected separately for each source for each of the four months in the reference period.

The edits for income amounts 1-56 are applied only to imputed amounts. No reported cross-sectional amounts are changed. If all monthly amounts for all reference periods for a specific income source were imputed, these imputed amounts are averaged and the average imputed amount replaces the original imputed amounts; otherwise, imputed amounts are replaced by reported amounts obtained

---

[19] The calculated amount is computed by multiplying the number of weeks with pay each for each month by: 1) the number of usual hours worked per week; and 2) the hourly wage rate.

[20] See Appendix 1 for a description of nonwage and salary income sources 1-56 and asset types 100-150.

[21] At the end of each interview the sources of income, but not amounts, are transcribed from the Core questionnaire to the Control card. Just before the next interview the sources of income are transcribed from the Control card to the current wave Core questionnaire, which is used as the source of information for reconciling and updating income sources.

from other reference periods. When both reported and imputed amounts are present on a record, the imputed amounts are replaced with the nearest reported amount. The implementation of the nearest neighbor concept gives priority to the first month with a reported values preceding the month containing an imputed value. The monthly income amount which meets this criterion replaces the imputed amount. The first succeeding month with a reported value is used as a replacement value only when no month prior to the month requiring replacement contains a reported amount.

This editing protocol for replacing wave-imputed values typically produces strings of equal monthly amounts with a value equal to the last reported amount because when an amount is imputed it almost always is imputed for each month in the reference period. Since most monthly amounts are reported in this manner (i.e., the amount reported for each month is the same) the editing procedure for income sources 1-56 replicates the most frequent reporting pattern. Table 4.3 contains rates of wave-imputed item missing data for selected monthly nonwage and salary income amounts from the 1984 SIPP panel which were subject to longitudinal editing. The table

Table 4.3 Rates of Longitudinal Editing for Selected Monthly Nonwage and Salary Income Amounts Missing in the Relevant Wave and Imputed: 1984 SIPP Longitudinal File, 32-Month Average *

Income Type	Total	Number of Records with Imputed Monthly Income Amount			Percent		
		None	Some	All	None	Some	All
Social security	6,422	5,630	550	242	87.7	8.6	3.8
Federal SSI	703	634	30	39	90.2	4.3	5.5
Unemployment compensation	494	428	22	44	86.6	4.5	8.9
Veterans compensation	687	597	53	37	86.9	7.7	5.4
AFDC	610	562	29	19	92.1	4.8	3.1
WIC	274	236	26	12	86.1	9.5	4.4
Food stamps	1,320	1,224	56	40	92.7	4.2	3.0
Child support	635	580	29	26	91.3	4.6	4.1
Company or union pension	1,615	1,363	130	122	84.4	8.0	7.6
Civil service pension	367	316	23	28	86.1	6.3	7.6
Military retirement	253	217	12	23	85.8	5.1	9.1
State/local government pension	598	512	42	44	85.6	7.0	7.4

* Includes imputations due to item nonresponse only; Type Z imputations are not included.

indicates the total number of recipients for the indicated income type, the number of recipients with no item missing data and the number of recipients with item missing data for some or all of the reference months.

Edits for Asset Types 100-150: the manner in which a sample person's asset profile is established and the way in which errors in asset ownership are detected for asset types 100-150 are nearly identical to the procedure used for income sources 1-56: the profile is initially established in the Wave 1 interview and reconciled and updated during each subsequent interview. Instead of using the nearest neighbor concept that is used for income types 1-56, any values for asset types 100-150 which were imputed during the cross-sectional edits were replaced with the average of the reported values from other waves.

The recording of income flows from asset types 100-150 on the questionnaire, however, is considerably different than for income sources 1-56. Although amounts are collected separately for each type of asset just as with income sources 1-56, other aspects of recording asset amounts are different. First, only the total asset amount is recorded for the reference period rather than individual monthly amounts. Second, because asset ownership can be held individually or jointly with two or more persons living in the same or different households, additional questions are required to clarify ownership patterns of and income flows from assets. Third, income flows from some assets are grouped and recorded as a total. For example, the separate income amounts from asset sources 100-103, 104-107 and 140-150 are summed and recorded as three single values. Although a single four-month total amount is recorded on the questionnaire for each asset source or group of asset sources, the wave and longitudinal record contain four equal monthly amounts for each reference period in which an asset had an income flow, which is derived by dividing the four-month total amount by four. Also, joint amounts received by husbands and wives are divided equally between the husband and wife so that amounts appear separately on each person's record even though the total amount received jointly was recorded on either the husband's or wife's questionnaire. Table 4.4 shows counts of edits that substituted average reported amounts for imputed data by asset type.

4.6 Longitudinal Edits to Eliminate Duplicate Reporting of AFDC, Food Stamps and WIC Income Amounts

The primary means of detecting duplicate reporting of income amounts for AFDC, Food stamps and WIC by both the husband and wife are through item checks in the questionnaire. Any additional instances of duplicate reporting are identified during longitudinal processing by locating husbands and wives reporting amounts for the same income source for the same month and deleting either the husband's or wife's amount.[22]

4.7 Longitudinal Editing of Program Coverage Variables

An important function of the SIPP questionnaire design is to identify the composition of specific "transfer units" within the household. A transfer unit is defined as a group of persons who qualify for and receive a cash or noncash benefit. Transfer units are identified by first determining the primary recipient. The primary recipient, in turn, identifies other household members who are included as part of the group qualifying for benefits. The primary recipient identifies other members of the transfer unit following questions which ask for the monthly amount of the benefit, except for Medicaid coverage which is recorded in the recipiency section of the questionnaire. Membership in a transfer unit relates to the entire reference period whether or not someone was eligible during each of the four months. Coverage indicators which identify whether a household member received a particular benefit are created during cross-sectional processing and are based on information provided by the primary recipient. The coverage indicators created during cross-sectional processing, however, do not identify members of a particular transfer unit specifically. The longitudinal editing

[22] NOTE: there are no corresponding recipiency items to be edited because the amounts are used as recipiency indicators in the longitudinal file.

system restructures coverage variables to allow users to identify members of transfer units. The monthly program coverage fields for the income/benefit types listed below are structured to allow identification of individual
program units within a particular household:

* Aid to families with dependent children;
* Food stamps;
* WIC;
* Veterans pensions and compensation;
* General assistance;
* Other welfare;
* Foster child care;
* Indian, Cuban and Refugee Assistance;
* Social security (children only); and
* Railroad retirement (children only).

In this procedure for identifying program units the person numbers of the household members covered were used to form the program units. The program units were numbered from 1 to N. All persons in the same program unit for a particular income/benefit type in a particular month are assigned the sequence number of the person's record for the person in whose name

Table 4.4 Rates of Longitudinal Editing for Asset Types 100-150: 1984 SIPP Longitudinal File, 32-Month Average *

Asset Type	Total	Number of Records with Imputed Monthly Income Amount			Percent		
		None	Some	All	None	Some	All
100-103, Joint	11,756	9,889	1,262	605	84.1	10.7	5.1
100-103, Own	9,880	7,720	1,247	913	78.1	12.6	9.2
104-107, Joint	1,124	916	175	33	81.5	15.6	2.9
104-107, Own	1,122	831	173	118	74.1	15.4	10.5
110, Joint, Received	1,173	965	89	119	82.3	7.6	10.1
110, Joint, Credited	586	319	114	153	54.4	19.5	26.1
110, Own, Received	2,144	1,105	287	752	51.5	13.4	35.1
110, Own, Credited	1,166	397	211	558	34.0	18.1	47.9
120, Joint	1,336	1,021	206	109	76.4	15.4	8.2
120, Received	512	359	80	73	70.1	15.6	14.3
120, Other, Joint	239	172	27	40	72.0	11.3	16.7
130, Joint	446	362	59	25	81.2	13.2	5.6
130, Own	258	162	45	45	62.8	17.4	17.4
140-150	595	505	44	46	84.9	7.4	7.7

* Includes imputations due to item nonresponse only; Type Z imputations are not included.

the program was reported. A value of zero in a program coverage variable indicates a "not covered" status. If, in the process of assigning the program unit identifiers, a person is listed as a member of more than one unit for the same income/benefit type, the unit identifier of the first unit identified during the processing of that household's data for that month is assigned. During the development of the transfer unit indicator it was revealed that respondents sometimes incorrectly report that "all" persons in the household are covered by a particular program. Most of these errors have been eliminated during longitudinal processing.

4.8 Longitudinal Edits for Health and Medical Care Coverage Variables

The private health insurance variables on the longitudinal file are structured as three variables: 1) a variable indicating coverage in the person's "own name" (variable name: "HIOWNCOV"); 2) a variable indicating coverage in "someone else's name" (variable name: "HIOTCOV"); and 3) a variable indicating if the insurance was obtained through an employer (variable name: "HIEMPLYR"). This last variable applies only to persons with coverage in their own name. Unlike the cross-sectional files which list person numbers of covered individuals on the records of the person in whose name the policy is held, no attempt was made to establish covered units; that is, which household members were covered by which member's policy.

The Medicaid coverage field on the longitudinal file also differs in structure from the field on the cross-sectional files. The detailed responses that are included on the cross-sectional files are not included on the longitudinal files. Only the "CAIDCOV" field which reflects the fully edited coverage indicator is included for each of the reference months.

5. Assessing the Influence of Imputed Data on Analyses

All surveys experience missing data to one degree or another. In the SIPP missing data occur when responding sample persons refuse to provide or are unable to provide requested information, provide imprecise or inaccurate information, interviewers forget to ask a question or incorrectly record a response, or a response is inconsistent with related responses or incompatible with response categories. Missing data create problems for analysts for a number of reasons. First, data sets which contain missing data are not as convenient to analyze as data sets which are complete. Second, consistency between analyses are not guaranteed in the presence of missing data because different analyses may be based on different subsets of the data depending on the pattern of missing data. Third, the bias component of the mean square error is increased in the presence of nonignorable nonresponse, which leads to biased estimates of population parameters. Whether analysts apply a specific mechanism for handling missing data in their analyses, such as imputation, all analyses of survey data make implicit or explicit assumptions about patterns of missing data. Analyses based on data sets which are not imputed for missing data implicitly assume that missing data are missing at random in the population at large. The imputation procedures used in the SIPP also make the assumption that missing data are missing at random, but the assumption is more tenable and explicit because the missing at random assumption is made within subgroups of the population rather than for the population at large.

The imputation procedures used in the SIPP are designed to be suitable for a number of analytical purposes rather than ideal for any one specific application. The preceding chapters discussed the various types of imputation procedures used in the SIPP to compensate for missing data. Other types of adjustment strategies applied to SIPP data include the large class of edits and the assignment of weights, a component of which adjusts for unit nonresponse. Although neither the editing or weighting procedures are reviewed in this report users of SIPP data should become familiar with these adjustment procedures. This final chapter reviews some general guidelines for assessing the effect of imputed data on analyses. The guidelines which are outlined below are limited to situations in which users can identify an item or record which has been statistically imputed, such as for individual SIPP wave files. In these instances imputation

Table 5.1 Percent Cumulative Nonresponse Rate by Wave for Selected SIPP Panels*

SIPP Panel	SIPP Wave							
	1	2	3	4	5	6	7	8
1984	4.9	9.4	12.3	15.4	17.4	19.4	21.0	22.0
1985	6.7	10.8	13.3	16.3	18.8	19.7	20.5	20.8
1986	7.3	13.4	15.2	17.1	19.3	20.0	20.6	--

* Source: Singh et al. (1990).

flags provide users with choices about how to proceed. Data whose values have been changed through wave or longitudinal editing protocols or deductive imputation, on the other hand, are not flagged and cannot be readily identified.

For users of SIPP data who are interested in assessing the influence of imputed data on their analyses, the first issue to address is whether SIPP imputation procedures have properties which meet their specific analytical requirements. If not, users have the option of reimputing the data using another procedure. Although reimputing the data will not be a practical alternative for many users, the process of evaluating the appropriateness of SIPP imputation procedures for particular analytical applications serves to inform the analyst about potential difficulties. A general discussion of the treatment of missing data in sample surveys is given by Kalton and Kasprzyk (1986). Sedransk (1985), Little (1986), and Jann-Huei and Sedransk (1987) discuss properties of commonly used imputation procedures. An example of the impact of imputation procedures on the distributional characteristics of a population of low income persons is discussed by Doyle and Dalrymple (1987).

An evaluation of the effect of imputed data on analyses should include a review of rates of unit nonresponse across waves and an assessment of the extent of item missing data. Table 5.1 contains rates of nonresponse for the 1984, 1985 and 1986 SIPP panels and shows that rates of nonresponse are quite low in the early waves of a SIPP panel and accumulate to around 20 percent over the life of a panel. The nature and extent of nonresponse affects imputation outcomes in two important ways. First, the likelihood that nonresponse is nonignorable increases as rates of nonresponse increase, particularly when the refusal component of nonresponse increases. All SIPP imputation procedures assume that nonresponse is ignorable within subgroups which define the imputation matrices; that is, that the nonresponding cases are a random subset of the responding cases within a subgroup. Second, as the percentage of cases reinterviewed drops over time the potential pool from which donors are selected shrinks. As the pool of potential donors decreases the possibility that donors are used more than once increases, which increases the variance of an estimate. The rates on nonresponse in Table 5.1 are for the total sample at each wave. Similar tables can be produced for subgroups which are important to one's analysis.

The level of item missing data is important to assess because imputation procedures were originally developed to handle small amounts of missing data. In general, the effect of imputation will be small for items with low rates of missing data (for estimates of means, totals, proportions and distributions, but not necessarily or relationships between variables). One needs to assess, however, whether rates of item missing data are high among important subclasses. Recall that the standard error of an estimate is inversely proportional to sample size; as the number of imputed values increases the effective sample size decreases which can result in substantial underestimation of variances unless the estimation algorithm explicitly accounts for increases in variance due to imputation. Rates of missing data for Type Z imputed cases is provided in Table 2.3. Rates of item missing data for selected variables, waves and panels are displayed in Table 3.5 and illustrates that the extent of missing data varies considerably across items. Similar tables should be prepared for variables and subgroups important for one's analysis. Lepkowski et al. (1987) provide a framework for evaluating the effect of imputed values on analyses using data from a large federal survey. Users of SIPP data can adapt this framework to their own analyses.

The availability of imputation flags on SIPP public use files, which indicate when a missing data value has been imputed, provide users with the option of conducting their analyses with and without imputed data. Comparing point estimates and their variances calculated with and without imputed data are important for all analyses. An imputation procedure may affect descriptive statistics such as means and totals in one way and complex statistics such as regression coefficients, variances and correlations in quite a different way. Santos (1981), for example, has shown that multivariate relationships based on nonimputed values can be significantly altered when imputed values are included in the analysis.

List of References

Coder, J.F. (1978) Income Data Collection and Processing from the March Income Supplement to the Current Population Survey. The Survey of Income and Program Participation Proceedings of the Workshop on Data Processing, February 23-24, 1978, (Eds. D. Kasprzyk), Chapter II. Washington, D.C.: U.S. Department of Health, Education and Welfare.

Doyle, P. and Dalrymple, R. (1987) The Impact of Imputation Procedures on Distributional Characteristics of the Low Income Population, Proceedings of the Bureau of the Census Third Annual Research Conference, Washington D.C., Department of Commerce, PP. 483-508.

Jabine, Thomas, B. (1990) Survey of Income and Program Participation Quality Profile, Second Edition, U.S. Bureau of the Census.

Jinn, Jann-Huei and Sedransk, J. (1987) Effect on Secondary Data Analysis of Different Imputation Methods, Proceedings of the Bureau of the Census Third Annual Research Conference, U.S. Department of Commerce, Washington D.C., pp. 509-530.

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

Kalton, G. and Kasprzyk, D. (1982) Imputing for Missing Survey Responses. Proceedings of the Section on Survey Research Methods, American Statistical Association. pp. 22-31.

Kalton, G. and Kasprzyk, D. (1986) The Treatment of Missing Survey Data. Survey Methodology, Vol. 12, No. 1, pp. 1-16.

Kalton, G., Kasprzyk, D. and Santos, R. (1981) Issues of Nonresponse and Imputation in the Survey of Income and Program Participation. In Current Topics in Survey Sampling, Proceedings of the International Symposium on Survey Sampling, (Eds. D. Krewski, R.
Platek, J.N.K. Rao), New York, Academic Press, pp. 455-480).

Kish, L. (1965) Survey Sampling, John Wiley & Sons, New York.

Lepkowski, J.M, Landis, R.L., and Stehouwer, S.A. (1987) Strategies for the Analysis of Imputed Data Froma Sample Survey, Medical Care, Vol, 25, No. 8., pp.705-716.

Little, J.A. Roderick (1986) Missing Data in Census Bureau Surveys, Proceedings of the Bureau of the Census Second Annual Research Conference, U.S. Department of Commerce, Washington D.C., pp. 442-454.

Nelson, D., McMillen, D. and Kasprzyk (1985) An Overview of the Survey of Income and Program Participation: Update 1, SIPP Working Paper Series No. 8401, U.S. Bureau of the Census., Washington D.C.

Sande, I.G. (1982) Imputation in Surveys: Coping with Reality. The American Statistician, Vol. 36, pp. 145-152.

Sande, I.G. (1983) Hot-Deck Imputation Procedures. In Incomplete Data in Sample Surveys, Vol. 3, Proceedings of the Symposium, (Eds. W.G. Madow and I. Olkin), New York: Academic Press, pp.339-349.

Santos, R.L. (1981) Effects of Imputation on Regression Coefficients, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 140-145.

Sedransk, J. (1985) The Objectives and Practice of Imputation, Proceedings of the Bureau of the Census First Annual Research Conference, Washington D.C., U. S. Department of Commerce, pp. 445-452.

Singh, R.P., Huggins, V., and Kasprzyk, D. (1990) Handling Single Wave Nonresponse In a Panel Survey, SIPP Working Paper Series No. 9009, U.S. Bureau of the Census, Washington D.C.

Singh, R.P. and Petroni, R.J. (1988) Nonresponse Adjustment Methods for Demographic Surveys at the U.S. Bureau of the Census, SIPP Working Paper Series No. 8823.

Survey of Income and Program Participation (SIPP) 1984 (and 1987) Full Panel Microdata Research File, Technical Documentation. U.S. Bureau of the Census, Washington, DC (1990).

U.S. Bureau of the Census Memoranda, SIPP 1984 Panel Cross-Sectional Imputation System Hot-Deck Matrices for Core Item Nonresponse (September 12, 1984).

U.S. Bureau of the Census Memoranda, Summary of Speech on Longitudinal Demographic Edits (December 16, 1986).

U.S. Bureau of the Census Memoranda, The 1984 SIPP Panel File (June 1, 1988).

U.S. Bureau of the Census Memoranda, Type Z Imputation Procedures for Waves 2 through 9 of the 1984 Panel of the SIPP (September 14, 1984).

Welniak, E.J. and Coder, J.F. (1980) A Measure of the Bias in the March CPS Earnings Imputation System. Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 421-425.

Appendix 1: Income and Asset Sources

Income Sources:

Code	Description
1	Social Security
2	Railroad Retirement
3	Federal Supplemental Security Income (SSI)
5	State unemployment compensation
6	Supplemental unemployment benefits
7	Other unemployment compensation
8	Veterans' compensation or benefits
10	Workers' compensation
12	Employer or union temporary sickness policy
13	Payments from a sickness, accident, or disability insurance policy purchased on own
20	Aid to families with dependent children (AFDC, ADC)
21	General assistance or general relief
23	Foster child care payments
24	Other welfare
25	Women, Infants, and Children (WIC)
27	Food stamps
28	Child support payments
29	Alimony payments
30	Pension from company or union
31	Federal civil service or other federal civilian employee pensions
32	U.S. military retirement pay
34	State government pensions
35	Local government pensions
36	Income from paid up life insurance policies or annuities
37	Estates and trusts
38	Other payments for retirement, disability or survivor
40	GI bill education benefits
41	Other VA educational assistance
50	Income assistance from a charitable group
51	Money from relatives or friends
52	Lump sum payments
53	Income from roomers or boarders
54	National guard or reserve pay
55	Incidental or casual earnings
56	Other cash income not included elsewhere
75	State SSI/black lung/state temporary disability benefits/indian, cuban or refugee assistance/national guard or reserve forces retirement

Asset Sources:

Code	Description
100	Regular or passbook savings accounts
101	Money market deposit accounts
102	Certificates of deposit or other savings certificates
103	NOW, Super NOW, or other interest-earning checking accounts
104	Money market funds
105	U.S. Government securities
106	Municipal or corporate bonds
130	Mortgages
174	U.S. Savings Bonds (E, EE)
107	Other interest-earning assets such as mutual bond funds, unit trusts, money loaned to a private individual, etc.
110	Stocks and mutual fund shares
120	Rental property
140	Royalties
150	Other financial investments such as investments in a non-corporate business venture managed by others, investments in a closely-held corporation, etc.

Appendix 2: Overview of Type Z Imputation Procedures

The imputation of Core questionnaire data for Type Z and Departure noninterviews is completed in three steps.* In the first step recipient and donor cases are identified. In the second step, each noninterview case is matched with four or five donor records depending on whether sample persons were interviewed in the previous wave. In the final step the one donor record which represents the best match is selected and duplicated for the noninterview record.

Step 1: Identify Person-Level Missing Data and Potential Donor Records

1.1 Identify persons for whom entire records will be imputed and whether they were interviewed in the previous wave.

1.2 Classify noninterviews into disjoint groups according to values on a set of matching variables. Create 4 Type A matching records for sample persons not interviewed in the previous wave and 5 Type B matching records for sample persons interviewed in the previous wave. Each matching record contains the following variables: 1) SUSEQNUM, the sequence number of the sample unit containing the person whose record will be imputed; 2) PP-RCSEQ, the relative position within the sample unit of the person whose record will be imputed; 3) the type of match: A or B; 4) the level of match:1-4 for persons not interviewed in the previous wave and 1-5 for persons interviewed in the previous wave; and, 5) a match index which represents the product of the values of each of the match variables.

1.3 Identify donor records and whether person was interviewed in previous wave.

1.4 Create comparable donor file containing 4 Type A matching records **and** 5 Type B matching records for persons interviewed in the previous wave and only 4 Type A matching records for persons not interviewed in the previous wave.

Step 2: Sort the Files

* Type Z noninterviews occur when a member of an interviewed household is not interviewed because they are unavailable for an interview or refuse and a proxy interview is not obtained. Departure noninterviews include persons who were members of a SIPP interviewed household sometime during the four-month reference period but were no longer a household member on the date of interview. The phrase "Departure Noninterview" is not an official Census term.

2.1 Sort both files on the three match keys: match type, level of match and match index.

EXAMPLE: This example shows a portion of a file containing 5 Type A and 3 Type B noninterview matching. The respondent file has 3 Type A records and 5 Type B records. The Type A and Type B records each have the same value of match level and match index.

SUSEQNUM	PP-RCSEQ	MATCH TYPE	LEVEL	INDEX
Noninterviews:				
9	1	A	3	17956
27	2	A	3	17956
54	2	A	3	17956
90	2	A	3	17956
206	1	A	3	17956
407	1	B	4	19789
489	3	B	4	19789
609	1	B	4	19789

SUSEQNUM	PP-RCSEQ	MATCH TYPE	LEVEL	INDEX
Respondents:				
12	1	A	3	17956
24	3	A	3	17956
53	3	A	3	17956
76	1	B	4	19789
154	1	B	4	19789
345	1	B	4	19789
431	2	B	4	19789
676	2	B	4	19789

Step 3: Match Both Files on Match Keys and Identify Best Match.

3.1 Match the two files on match type, level of match and match index. If the match is one to many, that is, the donor file for a match group[**] contains more records than the noninterview match file, match the first record in the sorted noninterview file with the first record in the sorted donor file. Continue to sequentially match noninterview records with donor records until all noninterview records have been processed in a match group. If the donor file contains fewer records than the noninterview file for a match group, match the records sequentially until all donor records with a match group have been used, at which point return to the beginning of the match group in the donor file and continue to sequentially match noninterview records with donor records until all records in the noninterview match group have been processed. When a match is found update the noninterview record with information from the donor record. If no match is found for a level, the donor fields on the noninterview record are updated with zeros, indicating no match was found.

[**] A match group is a set of records which have the same values on each match variable.

EXAMPLE:The result of the matching operation is a composite record for each noninterview record. Note that the Type A match in this example required more donors than were available. In this case additional donors are obtained sequentially from the top of the match group in the donor file.

Noninterview Values		Respondent Values		Values Common to Noninterview and Respondent Record		
SUSEQNUM	PP-RCSEQ	SUSEQNUM	PP-RCSEQ	MATCH TYPE	LEVEL	INDEX
9	1	12	1	A	3	17956
27	2	24	3	A	3	17956
54	2	53	3	A	3	17956
90	2	12	1	A	3	17956
206	1	24	3	A	3	17956
407	1	76	1	B	4	19789
489	3	154	1	B	4	19789
609	1	345	1	B	4	19789

3.2 Sort the matched file by the noninterview values for the variables SUSEQNUM, PP-RCSEQ, match type and match level. This sort brings back together all the matching records associated with a sample person whose information is to be imputed. Select the donor which corresponds to the lowest numbered match level found. This record constitutes the best match because it utilized more variables containing more detail than any other level of match.

EXAMPLE: For Case 9, no level 1 or 2 match was found. The best match was obtained at level 3; therefore; the first person in the twelfth sample unit will be used to impute the first person in the ninth sample unit. For Case 407 the best match was obtained at level one, so the first person in the 76th sample unit will be used to impute the first person in the 407th sample unit.

Noninterview Values		Respondent Values			
SUSEQNUM	PP-RCSEQ	SUSEQNUM	PP-RCSEQ	MATCH TYPE	LEVEL
Case 9:					
9	1	0	0	A	1
9	1	0	0	A	2
9	**1**	**12**	**1**	**A**	**3**
9	1	59	1	A	4
Case 407:					
407	**1**	**76**	**1**	**B**	**1**
407	1	264	1	B	2
407	1	769	1	B	3
407	1	904	1	B	4
407	1	951	1	B	5