MATHEMATICA
CENTER FOR
STUDYING DISABILITY POLICY

REPORT

FINAL REPORT

National Beneficiary Survey-General Waves Round 5 (Volume 1 of 3): Editing, Coding, Imputation, and Weighting Procedures

August 17, 2017

Eric Grau
Sara Skidmore
Yuhong Zheng
Hanzhi Zhou
Debra Wright
Kirsten Barrett

Submitted to:
Social Security Administration
Office of Research, Demonstrations, and Employment Support
500 E. St., SW, 9th Floor
Washington, DC 20254
Project Officer: Mark Trapani

Contract Number: 0600-12-60094

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763

Project Director: Jason Markesich

Reference Number: 40160.124

This page has been left blank for double-sided copying.

CONTENTS

ACRONYMS vii

NBS DATA DOCUMENTATION REPORTS ix

I INTRODUCTION 1

A. NBS–General Waves Objectives 1

B. NBS–General Waves Sample Design Overview 2

C. NBS–General Waves Round 5 Survey Overview 3

1. Completes and Response Rates 5

2. Nonresponse Bias 5

II DATA EDITING AND CODING 6

A. Data Editing 7

B. Coding Verbatim Responses 7

1. Coding Open-Ended, “Other/Specify,” and Field-Coded Responses 8

2. Health Condition Coding 9

3. Industry and Occupation 12

III SAMPLING WEIGHTS 14

A. Computing and Adjusting the Sampling Weights: A Summary 15

1. Quality Assurance 19

B. Details of Calculation of Weights 19

1. Base Weights 19

2. Response Rates and Nonresponse Adjustments to the Weights 20

3. Post-Stratification 31

IV IMPUTATIONS 33

A. NBS Imputations of Specific Variables 35

1. Section L: Race and Ethnicity 36

2. Section B: Disability Status Variables and Work Indicator 37

3. Section C: Current Jobs Variables 38

4. Section I: Health Status Variables 40

5. Section K: Sources of Income Other Than Employment 44

6. Section L: Personal and Household Characteristics 45

V ESTIMATING SAMPLING VARIANCE 47

REFERENCES 49

APPENDIX A: OTHER SPECIFY AND OPEN-ENDED ITEMS WITH ADDITIONAL CATEGORIES CREATED DURING CODING

APPENDIX B: SOC MAJOR AND MINOR OCCUPATION CLASSIFICATIONS

APPENDIX C: NAICS INDUSTRY CODES

APPENDIX D: PARAMETER ESTIMATES AND STANDARD ERRORS FOR NONRESPONSE MODELS

APPENDIX E: SUDAAN PARAMETERS FOR NATIONAL ESTIMATES FROM THE NBS-GENERAL WAVES ROUND 5 SAMPLE

TABLES

I.1 NBS–General Waves Round 5 Actual Sample Sizes, Target Completes, and Completes 3

I.2 Sources of Error, Description, and Methods to Minimize Impact 3

II.1 Supplemental Codes for “Other/Specify” Coding 9

II.2 ICD-9 Category and Supplemental Codes 10

II.3 Supplemental Codes for Occupation and Industry Coding 13

III.1 Study Population (as of June 30, 2014), Initial Augmented Sample Sizes, and Initial Weights by Sampling Strata in the National Beneficiary Survey 19

III.2 Weighted Location, Cooperation, and Response Rates for Representative Beneficiary Sample, by Selected Characteristics 22

III.3 Location Logistic Propensity Model: Representative Beneficiary Sample 28

III.4 Cooperation Logistic Propensity Model: Representative Beneficiary Sample 28

IV.1 Race and Ethnicity Imputations 36

IV.2 Disability Status Imputations 38

IV.3 Current Jobs Imputations 40

IV.4 Health Status Imputations, Questionnaire Variables 41

IV.5 Health Status Imputations, Constructed Variables 43

IV.6 Imputations on Sources of Income Other Than Employment 44

IV.7 Imputations of Personal and Household Characteristics 46

This page has been left blank for double-sided copying.

ACRONYMS

AIC	Akaike's Information Criterion
CAPI	Computer-assisted personal interviewing
CATI	Computer-assisted telephone interviewing
CHAID	Chi-Squared Automatic Interaction Detector
ICD-9	International Classification of Diseases, 9th Revision
MSA	Metropolitan statistical area
NAICS	North American Industry Classification System
NBS	National Beneficiary Survey
PSU	Primary sampling unit
RBS	Representative Beneficiary Sample
SAS	Statistical software, formerly Statistical Analysis System (SAS is a registered trademark of SAS Institute Inc., of Cary, North Carolina)
SGA	Substantial Gainful Activity
SOC	Standard Occupational Classification
SPSS	Statistical Package for the Social Sciences (SPSS is a registered trademark of SPSS Inc., of Chicago, Illinois)
SSA	Social Security Administration
SSDI	Social Security Disability Insurance (Title II of the Social Security Act)
SSI	Supplemental Security Income (Title XVI of the Social Security Act)
SSU	Secondary sampling unit
STATA	Statistical software (STATA is a registered trademark of StataCorp LP, of College Station, Texas)
SWS	Successful Worker Sample
TRS	Telecommunications relay service
TTW	Ticket to Work and Self-Sufficiency

This page has been left blank for double-sided copying.

NBS DATA DOCUMENTATION REPORTS

The following publically available reports are available from SSA on their website (https://www.ssa.gov/disabilityresearch/nbs_round_5.htm#general):

- **User's Guide for Restricted- and Public-Use Data Files** (Wright et al. 2017). This report provides users with information about the restricted-use and public-use data files, including construction of the files; weight specification and variance estimation; masking procedures employed in the creation of the Public-Use File; and a detailed overview of the questionnaire design, sampling, and NBS–General Waves data collection. The report provides information covered in the Editing, Coding, Imputation and Weighting Report and the Cleaning and Identification of Data Problems Report (described below) —including, procedures for data editing, coding of open-ended responses, and variable construction—as well as a description of the imputation and weighting procedures and development of standard errors for the survey. In addition, this report contains an appendix addressing total survey error and the NBS.
- **NBS Public-Use File codebook** (Bush et al. 2017). This codebook provides extensive documentation for each variable in the file, including variable name, label, position, variable type and format, question universe, question text, number of cases eligible to receive each item, constructed variable specifications, and user notes for variables on the public-use file. The codebook also includes frequency distributions and means as appropriate.
- **NBS–General Waves Questionnaire** (Barrett et al. 2016). This document contains all items on Round 5 of the NBS–General Waves and includes documentation of skip patterns, question universe specifications, text fills, interviewer directives, and checks for consistency and range.
- **Editing, Coding, Imputation, and Weighting Report** (current report). In this report, we summarize the editing, coding, imputation, and weighting procedures as well as the development of standard errors for Round 5 of the NBS–General Waves. It includes an overview of the variable naming, coding, and construction conventions used in the data files and accompanying codebooks; describes how the sampling weights were computed to the final post-stratified analysis weights for the representative beneficiary sample; outlines the procedures used to impute missing responses; and discusses procedures that should be used to estimate sampling variances for the NBS.
- **Cleaning and Identification of Data Problems Report** (Skidmore et al. 2017). This report describes the data processing procedures performed for Round 5 of the NBS–General Waves. It outlines the data coding and cleaning procedures and describes data problems, their origins, and the corrections implemented to create the final data file. The report describes data issues by sections of the interview and concludes with a summary of types of problems encountered and general recommendations.
- **NBS Nonresponse Bias Analysis** (Grau et al. 2017). The purpose of this report was to determine whether the nonresponse adjustments applied to the sampling weights of Round 5 of the NBS-General Waves appropriately accounted for differences between respondents and nonrespondents or whether the potential for nonresponse bias still existed.

The following restricted use report is available from SSA through a formal data sharing agreement:

- **NBS Restricted-Access Codebook** (Bush et al. 2017). In this codebook, we provide extensive documentation for each variable in the file, including variable name, label, position, variable type and format, question universe, question text, number of cases eligible to receive each item, constructed variable specifications, and user notes for variables on the restricted-access file. The codebook also includes frequency distributions and means as appropriate.

I. INTRODUCTION

As part of an evaluation of the National Beneficiary Survey–General Waves (NBS–General Waves) project, Mathematica Policy Research conducted the first of three rounds of data collection in 2015, with two additional rounds to be administered in 2017 and 2019. Sponsored by the Social Security Administration's (SSA) Office of Retirement and Disability Policy, the survey collected data from a national sample of SSA disability beneficiaries. Mathematica collected the data by using computer-assisted telephone interviewing (CATI). We used computer-assisted personal interviewing (CAPI) for follow-ups of CATI nonrespondents and for those who preferred or needed an in-person interview to accommodate their disabilities.

The prior rounds of the NBS—conducted by SSA in 2004, 2005, 2006, and 2010[1]—took an important first step toward understanding the work interest and experiences of Supplemental Security Income (SSI) recipients and Social Security Disability Insurance (SSDI) beneficiaries. These surveys helped glean information about beneficiaries' impairments; health; living arrangements; family structure; occupation before disability; and use of non-SSA programs (for example, the Supplemental Nutrition Assistance Program, or SNAP). The prior NBS rounds also evaluated the Ticket to Work and Self-Sufficiency (TTW) program. The NBS–General Waves no longer includes a focus on TTW. Instead, the survey seeks to uncover important information about the factors that promote beneficiary self-sufficiency and, conversely, the factors that impede beneficiary efforts to maintain employment.

In this report, we document the editing, coding, imputation, and weighting procedures, as well as the development of standard errors, for Round 5 of the NBS–General Waves. In Chapter II, we provide an overview of the variable naming, coding, and construction conventions that were used in the data files and accompanying codebooks. In Chapter III, we discuss how the initial sampling weights were computed to the final post-stratified analysis weight for the representative beneficiary sample. In Chapter IV, we describe the procedures used to impute missing responses for selected questions and in Chapter V we explain the procedures that should be used to estimate sampling variances for the NBS–General Waves. In Appendix A, we list the open-ended items that were assigned additional categories, as discussed in Chapter II. In Appendices B and C, we list the occupation and industry codes, respectively, which are also discussed in Chapter II. In Appendix D, we provide detailed parameter estimates and standard errors for the weight adjustment models, as discussed in Chapter III. Finally, in Appendix E, we present SUDAAN and SAS parameters for the national estimates from the Round 5 sample.

A. NBS–General Waves Objectives

The NBS–General Waves collects important beneficiary data that are not available from SSA administrative data or other sources. The survey addresses five major questions:

1. What are the work-related goals and activities of SSI and SSDI beneficiaries, particularly as they relate to long-term employment?

[1] In this report, we refer to the NBS rounds conducted in 2004, 2005, 2006, 2010, and 2015 as Round 1, Round 2, Round 3, Round 4, and Round 5, respectively. We refer to the planned 2017 and 2019 rounds as Round 6 and Round 7, respectively.

2. What are the short-term and long-term employment outcomes for SSI and SSDI beneficiaries who work?
3. What supports help SSA beneficiaries with disabilities find and keep jobs and what barriers to work do they encounter?
4. What are the characteristics and experiences of beneficiaries who work?
5. What health-related factors, job-related factors, and personal circumstances hinder or promote employment and self-sufficiency?

The NBS–General Waves captures information on SSA beneficiaries, including their disabilities, interest in work, use of services, and employment. SSA will combine data from Round 5 of the NBS–General Waves with SSA administrative data to provide critical information on access to jobs and employment outcomes for beneficiaries. As a result, SSA and external researchers who are interested in disability and employment issues may use the survey data for other policymaking and program planning efforts.

B. NBS–General Waves Sample Design Overview

During Round 5 of the NBS–General Waves, we fielded a nationally representative sample of 7,682 SSA disability beneficiaries (hereafter referred to as the representative beneficiary sample). Except for the stratification of the primary sampling units (PSUs), the sample design for the representative beneficiary sample (RBS) was nearly identical to the design of the RBS in Rounds 1 through 4.[2] The target population for the RBS consisted of SSI recipients and SSDI beneficiaries between the ages of 18 and full retirement age who resided in all 50 states and the District of Columbia, excluding outlying territories, and who were in an active pay status as of June 30, 2014.[3] As of that date, the target population consisted of approximately 13.8 million beneficiaries. We stratified the cross-sectional RBS by four age-based strata within the PSUs: (1) 18- to 29-year-olds, (2) 30- to 39-year-olds, (3) 40- to 49-year-olds, and (4) 50-year-olds and older. To ensure a sufficient number of persons seeking work, beneficiaries in the first three cohorts were oversampled (18- to 49-year-olds). The target number of completed interviews for Round 5 was 1,111 beneficiaries in each of the three younger age groups. For those 50 years and older, the target number of completed interviews was 667 beneficiaries. We summarize the actual sample sizes and number of completed interviews for both samples under the revised design in Table I.1.

For Round 5 of the NBS–General Waves we used a multistage sampling design. Because the geographical distribution of beneficiaries changed little between 2003 and 2011, we used the same 1,330 PSUs—which consist of one or more counties—that were created prior to Round 1.

[2] The Round 4 sample design included two samples, one for all beneficiaries (the RBS) and one for the ticket participants (the Ticket Participant sample). To accommodate the rollout of the ticket-to-work program, the primary sampling units (PSUs) were sampled within strata defined by the three phases of the rollout. The sample design for this round only includes one sample, that of all beneficiaries. The PSUs were not drawn within strata, except those defined by the two certainty PSUs.

[3] Active status includes beneficiaries who are currently receiving cash benefits as well as those whose benefits have been temporarily suspended for work or other reasons. Active status does not include beneficiaries whose benefits have been terminated.

The measure of size for each PSU in this sample was based upon the most current counts of beneficiaries. We selected a stratified national sample of 79 PSUs, with probability proportional to size.

Table I.1. NBS–General Waves Round 5 Actual Sample Sizes, Target Completes, and Completes

Sampling Strata	Sample Size	Target Completed Interviews	Actual Completed Interviews
Representative beneficiary sample	7,682	4,000	4,062
18- to 29-year-olds	2,268	1,111	1,149
30- to 39-year-olds	2,126	1,111	1,097
40- to 49-year-olds	2,076	1,111	1,104
50-year-olds or older	1,212	667	712

Source: NBS–General Waves Round 5.

C. NBS–General Waves Round 5 Survey Overview

The NBS was designed and implemented to maximize both response and data quality. Table I.2 describes the most significant sources of potential error identified at the outset of the NBS and describes the ways we attempted to minimize the impact of each. A more detailed discussion of our approach to minimizing total survey error can be found in Appendix A of the Round 5 User's Guide (Bush et al. 2017).

Table I.2. Sources of Error, Description, and Methods to Minimize Impact

Sources of Error	Description	Methods to Minimize Impact
Sampling	Error that results when characteristics of the selected sample deviates from the characteristics of the population.	Select a large sample size; select primary sampling units with probability proportional to size, basing the measure of size for each PSU on the counts of beneficiaries in the study population; use stratified sampling by age categories to create units within each stratum as similar as possible.
Specification	An error occurring when the concept intended to be measured by the question is not the same as the concept the respondent ascribes to the question.	Cognitive interviewing during survey development[a] and pre-testing; use of proxy, if sample member is unable to respond due to cognitive disability
Unit nonresponse	An error occurring when a selected sample member is unwilling or unable to participate (failure to interview). This can result in increased variance and potential for bias in estimates if nonresponders have different characteristics than responders.	Interviewer training; intensive locating, including field locating; in-person data collection; refusal conversion; incentives; nonresponse adjustment to weights

Sources of Error	Description	Methods to Minimize Impact
Item nonresponse	An error occurring when items are left blank or the respondent reports that he or she does not know the answer or refuses to provide an answer (failure to obtain and record data for all items). This can result in increased variance and potential bias in estimates if nonresponders have different characteristics than responders.	Use of probes; allowing for variations in reporting units; assurance of confidentiality; assistance during interview; use of proxy, if sample member unable to respond due to cognitive disability; imputation on key variables
Measurement error	An error occurring as a result of the respondent or interviewer providing incorrect information (either intentionally or unintentionally). This may result from inherent differences in interview mode.	Same instrument used in both interview modes; use of probes; adaptive equipment; interviewer training, validation of field interviews; assistance during interview; use of proxy, if sample member unable to respond due to cognitive disability
Data processing errors	An error occurring in data entry, coding, weighting, or analysis.	Coder training; monitoring and quality control checks of coders; quality assurance review of all weighting and imputation procedures

[a]Conducted during survey development phase under a separate contract held by Westat.

We did not expect item nonresponse to be a large source of error because there were few obviously sensitive items. In fact, item nonresponse was greater than 5 percent only for select items asking for wages and household income. Unit nonresponse was the greater concern given the population, thus the survey was designed to be executed as a dual-mode survey. Mathematica made all initial attempts to interview beneficiaries using CATI. We sought a proxy respondent when a sample person was unable to participate in the survey because of his or her disability. To promote response among Hispanics, Mathematica provided the questionnaire in Spanish. For languages other than English or Spanish, interpreters, if available in the sample person's home, conducted interviews. We made a number of additional accommodations for those sample members with hearing or speech impairments, including using a telecommunications relay service (TRS) and amplifiers.

If Mathematica could not locate and contact a sample member by telephone, a field locator was deployed to make contact in person. Once located, the field locator attempted to facilitate an interview with the sample member via CATI, using a staff cell phone to call into the data collection center (or the sample member's own phone, if preferred). If a sample member could not complete the interview by telephone in this manner due to his or her disability, trained field staff conducted the interview in person using CAPI. To reduce measurement error, the survey instrument was identical in each mode.

We began Round 5 CATI data collection for the NBS in February 2015. In June 2015, Mathematica began in-person locating and CAPI, which continued concurrent with CATI through October 2015. As mentioned earlier, the NBS–General Waves Round 5 sample comprised 7,682 cases.

1. Completes and Response Rates

In total, Mathematica completed 4,062 interviews (including 40 partially completed interviews).[4] Of these, we completed 3,649 by CATI and 413 by CAPI. We deemed an additional 297 beneficiaries as ineligible for the survey.[5]

During Round 5, we completed proxy interviews with 771 sample members (19 percent of all completed interviews). Of the completed proxy interviews, approximately 60 percent needed a proxy because the caregiver deemed the sample member unable to respond due to an intellectual disability; 32 percent needed a proxy because the sample member failed the cognitive assessment.[6] The remaining 8 percent needed a proxy because they were unable to complete the interview, as they did not understand either the questions or the question-response sequence after passing the cognitive assessment. There were an additional 136 cases in which sample members could not participate in the interview and proxies could not be identified to complete it on their behalf. Of these cases, 112 (82 percent) were situations in which a gatekeeper reported an intellectual disability and could not serve as a proxy. The remaining 24 (18 percent) were cases in which sample members could not participate because they were unable to successfully complete the cognitive screener and could not identify a proxy to complete the interview.

The weighted response rate for the representative beneficiary sample was 62.6 percent. More information about sample selection and sampling weights is available in Grau et al. (2017).

2. Nonresponse Bias

Because the weighted response rates within the age strata ranged from 54.7 to 62.6 percent and the overall response rate was less than 80 percent, we conducted a nonresponse bias analysis at the conclusion of data collection using all 7,682 sample cases, to determine if there were systematic differences between respondents and nonrespondents that could result in nonresponse bias. In sum, our analysis indicates that differences did exist between responders and nonresponders among variables that were not controlled for in the sample design. However, the nonresponse adjustments to the weights alleviated all known differences observed in the beneficiary sample. Some estimates from respondents using nonresponse-adjusted weights differed from the values in the sampling frame, but these mirrored differences that existed between the sampling frame and the entire sample using the initial sampling weights. The full nonresponse bias analysis can be obtained from SSA (https://www.ssa.gov/disabilityresearch/nbs_round_5.htm#general).

[4] Partial interviews were considered as completed if responses were provided through Section G of the interview.

[5] Ineligible sample members included those who were deceased, incarcerated, or no longer living in the continental United States and those whose benefit status was pending.

[6] The cognitive assessment was developed under a separate contract held by Westat.

This page has been left blank for double-sided copying.

II. DATA EDITING AND CODING

Prior to imputation, we edited and coded the NBS data to create the NBS data file. In this chapter, we document the variable naming, coding, and construction conventions that were used in the data files and accompanying codebooks.

A. Data Editing

At the start of data cleaning, we conducted a systematic review of the frequency counts of individual questionnaire items. We reviewed frequency counts by each questionnaire path to identify possible errors in skip patterns. We also reviewed interviewer notes and comments in order to flag and correct individual cases. As in earlier rounds, we edited only those cases that had an obvious data entry or respondent error. As a result, even though we devoted considerable time to a meticulous review of individual responses, we acknowledge that some suspect values remain on the file. (See Skidmore et al. [2017] for more detail on the editing and cleaning procedures.)

For all items with fixed field numeric responses (such as number of weeks, number of jobs, and dollar amounts), we reviewed the upper and lower values assigned by interviewers. Although data entry ranges were set in the CATI instrument to prevent the entry of improbable responses, the ranges were set to accommodate a wide spectrum of values in order to account for the diversity expected in the population of interest and to permit the interview to continue in most situations. For these reasons, we set extremely high and low values to missing (.D = don't know) in the case of apparent data entry error.

We included several consistency edit checks to flag potential problems during the interview. To minimize respondent burden, however, all consistency edit checks were suppressible. Although the interviewer was instructed to probe inconsistent responses, the interviewer could continue beyond a particular item if the respondent could not resolve the problem. In the post-interview stage, we manually reviewed remaining consistency problems to determine whether the responses were plausible. After investigating such cases, we either corrected them or set them to missing when we encountered an obvious error.

During data processing, we created several constructed variables to combine data across items. For these items, both the survey team and the analysis team reviewed the specifications. Several reviewers checked the SAS programming code. Finally, we reviewed all data values for the constructed variables based on the composite variable responses and frequencies.

For open-ended items assigned numeric codes, we examined frequencies to ensure the assignment of valid values. For health condition coding, we examined the codes to verify that the same codes were not assigned to both main and secondary conditions. Cases coded incorrectly were recoded according to the original verbatim response.

B. Coding Verbatim Responses

The NBS includes several questions designed to elicit open-ended responses. To make it easier to analyze the data connected with these responses, we grouped the responses and

assigned them numeric codes when possible. The methodology used to code each variable depended upon the variable's content.

1. Coding Open-Ended, "Other/Specify," and Field-Coded Responses

Three types of questions (described below) in the NBS did not have designated response categories; rather, the responses to the questions were recorded verbatim:

1. **Open-ended questions** have no response options specified. For example, Item G61 asks, "Why {were you/was NAME} unable to get these services?" For such items, interviewers recorded the verbatim response. Using common responses, we developed categories and reviewed them with analysts. Coders then attempted to code the verbatim response into an established category. If the response did not fit into one of the categories, coders coded it as "other."
2. **"Other/specify"** is a response option for questions with a finite number of possible answers that may not necessarily capture all possible responses. For example, "Did you do anything else to look for work in the last four weeks that I didn't mention?" For such questions, respondents were asked to specify an answer to "Anything else?" or "Anyone else?"
3. **Field-coded responses** are answers coded by interviewers into a predefined response category without reading the categories aloud to the respondent. If none of the response options seemed to apply, interviewers selected an "other/specify" category and typed in the response.

Based on an initial review of the data, we examined as part of data processing a portion of all verbatim responses in an attempt to uncover dominant themes for each question. We developed a list of categories and decision rules for coding verbatim responses to open-ended items. We also added supplemental response categories to some field-coded or "other/specify" items to facilitate coding if there were enough such responses and they could not be back-coded into pre-existing categories. (A list of all open-ended items that were assigned additional categories during the coding process appears in Appendix A.) Thus, we categorized verbatim responses for quantitative analyses by coding responses that clustered together (for open-ended and "other/specify" responses) or by back-coding responses into existing response options if appropriate (for field-coded and "other/specify" items). We applied categories developed during prior rounds of the NBS. In some cases, we added to the questionnaire categories developed in earlier rounds in order to minimize back-coding.

If the need for changes to the coding scheme became apparent during coding—for example, the addition of categories or clarification of coding decisions—we discussed and documented new decision rules. We sorted verbatim responses alphabetically by item for coders. The responses then lent themselves to filtering by coding status so that new decision rules could be easily applied to previously coded cases. When it was impossible to code a response, when a response was invalid, or when a response could not be coded into a given category, we assigned a two-digit supplemental code to the response (Table II.1). The data files exclude the verbatim responses. (See Skidmore et al. [2017] for full details on back-coding procedures.)

Table II.1. Supplemental Codes for "Other/Specify" Coding

Code	Label	Description
94	Invalid response	Indicates that this response should not be counted as an "other" response and should be deleted
95	Refused	Used only if verbatim response indicates that respondent refused to answer the question
96	Duplicate response	Indicates that the verbatim response already has been selected in a "code all that apply" item
98	Don't know	Used only if the verbatim response indicates that the respondent does not know the answer
99	Not codeable	Indicates that a code cannot be assigned based on the verbatim response

Source: NBS–General Waves Round 5.

2. Health Condition Coding

In Section B of the questionnaire, we asked each respondent to cite the primary and secondary physical or mental conditions that limit the kind or amount of work or daily activities that the respondent performs. Respondents could report main conditions in one of four questions: B2 (primary reason limited), B6 (primary reason eligible for benefits), B12 (primary reason formerly eligible for benefits if not currently eligible), and B15 (primary reason limited when first receiving disability benefits). The main purpose of the other items (B6, B12, and B15) was to collect information on a health condition from people who reported no limiting conditions in Item B2. For example, if respondents reported no limiting conditions, we asked if they were currently receiving Social Security benefits. If they answered "yes," we asked for the main reason that made them eligible for benefits (Item B6). If respondents said that they were not currently receiving benefits, we asked whether they had received disability benefits in the last five years. If they answered "yes," we asked for the condition that made them eligible for Social Security benefits (Item B12) or for the reason that first made them eligible if they no longer had that condition (Item B15). Respondents who said that they had not received disability benefits in the last five years were screened out of the survey and coded as ineligible. We assigned a value for the three health condition constructs to each response to Items B2, B6, B12, and B15. Although we asked respondents to cite one main condition in Items B2, B6, B12, or B15, many listed more than one. We maintained the additional responses under the primary condition variable and coded them in the order in which they were recorded.

For each item on a main condition, we asked respondents to list any other, or secondary, conditions. For example, in Item B4, we asked respondents who had reported a main condition in Item B2 to list other conditions that limited the kind or amount of work or daily activities they could perform. In Item B8, we asked respondents who had reported the main reason for their eligibility for disability benefits in Item B6 to list other conditions that made them eligible. For respondents who reported that they were not currently receiving benefits but who reported a main condition in Item B12 (the condition that made them eligible to receive disability benefits in the last five years), we asked in Item B14 for other reasons that made them eligible for benefits. For those who reported that their current main condition was not the condition that made them eligible for benefits and who were asked for the main reason for their initial

limitation, we also asked if any other conditions had limited them when they started receiving benefits (Item B17).

We coded respondents' verbatim responses by using the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9) five-digit coding scheme.[7] The ICD-9 is a classification of morbidity and mortality information developed in 1950 to index hospital records by disease for data storage and retrieval. The ICD-9 was available in hard copy for each coder. The coders, including many who had medical coding experience, attended an eight-hour training session before coding and were instructed to code to the highest level of specificity possible. We coded responses that were not specific enough for a five-digit code to four digits (subcategory) or three digits (category codes). We coded responses that were not specific enough for even three- or four-digit ICD-9 codes either as a physical problem (not specified) or to broader categories representing disease groups. In Table II.2, we list the broad categorical and supplementary codes. For cases in which the respondent reported several distinct conditions, all conditions were coded (for instance, three distinct conditions would be recorded and coded as B2_1, B2_2, and B2_3).

Table II.2. ICD-9 Category and Supplemental Codes

Code	Label	Description of ICD-9 Codes	Corresponding ICD-9 Codes
00	Other	Other and unspecified infectious and parasitic disease; alcohol dependence syndrome and drug dependence; learning disorders and developmental speech or language disorders; complications of medical care, not elsewhere classified (NEC)	136.0–136.9, 303.00–304.90, 315.00–315.39, 999.0–-999.9
01	Infectious and parasitic diseases	Borne by a bacterium or parasite and viruses that can be passed from one human to another or from an animal/insect to a human, including tuberculosis, HIV, other viral diseases, and venereal diseases (excluding other and unspecified infectious and parasitic diseases)	001.0–135, 137.0–139.8
02	Neoplasms	New abnormal growth of tissue (i.e., tumors and cancer), including malignant neoplasms, carcinoma in situ, and neoplasm of uncertain behavior	140.0–239.9
03	Endocrine/nutritional disorders	Thyroid disorders, diabetes, abnormal growth disorders, nutritional disorders, and other metabolic and immunity disorders	240.0–279.9
04	Blood/blood-forming diseases	Diseases of blood cells and spleen	280.0–289.9
05	Mental disorders	Psychoses, neurotic and personality disorders, and other nonpsychotic mental disorders, including mental retardation but excluding alcohol and drug dependence and learning, developmental, speech, or language disorders	290.0–302.9, 305.00–314.9, 315.4–319

[7] Although the ICD-10 was available at the time of coding, we used ICD-9 to be consistent with how we coded in previous rounds. More information on comparing ICD-9 codes to ICD-10 codes is available at http://www.qualityindicators.ahrq.gov/resources/Toolkits.aspx.

TABLE II.2 (continued)

Code	Label	Description of ICD-9 Codes	Corresponding ICD-9 Codes
06	Diseases of nervous system	Disorders of brain, spinal cord, central nervous system, peripheral nervous system, and senses, including paralytic syndromes and disorders of eye and ear	320.0–-389.9
07	Diseases of circulatory system	Heart disease; disorders of circulation; and diseases of arteries, veins, and capillaries	390-459.9
08	Diseases of respiratory system	Disorders of the nasal, sinus, upper respiratory tract, and lungs, including chronic obstructive pulmonary disease	460-519.9
09	Diseases of digestive system	Diseases of the oral cavity, stomach, esophagus, and duodenum	520.0-579.9
10	Diseases of genitourinary system	Diseases of the kidneys, urinary system, genital organs, and breasts	580.0-629.9
11	Complications of pregnancy, child birth, and puerperium	Complications related to pregnancy or delivery and complications of puerperium	630-677
12	Diseases of skin/ subcutaneous tissue	Infections of the skin, inflammatory conditions, and other skin diseases	680.0-709.9
13	Diseases of musculoskeletal system	Muscle, bone, and joint problems, including arthropathies, dorsopathies, rheumatism, osteopathies, and acquired musculoskeletal deformities	710.0-739.9
14	Congenital anomalies	Problems arising from abnormal fetal development, including birth defects and genetic abnormalities	740.0-759.9
15	Conditions in the perinatal period	Conditions that have origins in birth period, even if disorder emerges later	760.0-779.9
16	Symptoms, signs, and ill-defined conditions	Ill-defined conditions and symptoms; used when no more specific diagnosis can be made	780.01-799.9
17	Injury and poisoning	Problems that result from accidents and injuries, including fractures, brain injury, and burns (excluding complications of medical care NEC)	800.00–998.9
18	Physical problem, NEC	The condition is physical, but no more specific code can be assigned	No ICD-9 codes
95	Refused	Verbatim indicates that respondent refused to answer the question	No ICD-9 codes
96	Duplicate condition reported	The condition has already been coded for the respondent	No ICD-9 codes
97	No condition reported	The verbatim does not contain condition or symptom to code	No ICD-9 codes
98	Don’t know	The respondent reports that he or she does not know the condition	No ICD-9 codes
99	Uncodeable	A code cannot be assigned based on the verbatim response	No ICD-9 codes

Source: NBS–General Waves Round 5.

We employed several means to ensure that responses were coded according to the proper protocols. We performed an initial quality assurance check, per coder, for the first several cases that were coded. In addition, during coding, 10 percent of responses were randomly selected for review. In total, a supervisor reviewed approximately 20 percent of all coded responses, including cases flagged by coders for review because the coders were either unable to code them or did not know how to code them. Approximately 2 percent of all cases were recoded. In the course of the various reviews, we developed additional decision rules to clarify and document the coding protocol. We discussed the decision rules with coders and shared them to ensure that responses were coded consistently and accurately throughout the coding process. As for other open-ended items, when new decision rules were added, we reviewed previously coded responses and recoded them if necessary. After completion of the ICD-9 coding, we processed the health condition variables into a series of constructed variables that grouped health conditions into broad disease groups.

3. Industry and Occupation

In Section C of the questionnaire, we collected information about a sample member's current employment. In Section D of the questionnaire, we collected information about a sample member's employment in 2014. For each job, respondents were asked to report their occupation (Items C2 and D4) and the type of business or industry (Items C3 and D5) in which they were employed. To maintain comparability with earlier rounds, we used the Bureau of Labor Statistics 2000 Standard Occupational Classification (SOC) to code verbatim responses to these items.[8] The SOC classifies all occupations in the economy, including private, public, and military occupations, in which work is performed for pay or profit. Occupations are classified on the basis of work performed, skills, education, training, and credentials. The sample member's occupation was assigned one occupation code. The first two digits of the SOC codes classify the occupation to a major group and the third digit to a minor group. For the NBS–General Waves, we assigned three-digit SOC codes to describe the major group that the occupation belonged to and the minor groups within that classification (using the 23 major groups and 96 minor groups). We list the three-digit minor groups that are classified within major groups in Appendix B.

To maintain comparability with earlier rounds, we coded verbatim responses to the industry items according to the 2002 North American Industry Classification System (NAICS).[9] The NAICS is an industry classification system that groups establishments into categories on the basis of activities in which those establishments are primarily engaged. It uses a hierarchical coding system to classify all economic activity into 20 industry sectors. For the NBS–General Waves, we coded NAICS industries to three digits with the first two numbers specifying the industry sector and the third specifying the subsector. (Appendix C lists the broad industry sectors.) Most federal surveys use both the SOC and NAICS coding schemes, thus providing uniformity and comparability across data sources. Although both classification systems allow coding to high levels of specificity, SSA and the analysts decided based on research needs to limit coding to three digits.

[8] For more information, see *Standard Occupational Classification Manual, 2000,* or http://www.bls.gov/soc.

[9] For more information, see North American Industry Classification System, 2002, or http://www.naics.com/who-we-are/.

Mathematica developed supplemental codes for responses to questions about occupation and industry that could not be coded to a three-digit SOC or NAICS code (Table II.3). As we did in the health condition coding, we performed an initial quality assurance check, per coder, for the first several cases coded. Then, during coding, we randomly selected 10 percent of responses for review. In total, a supervisor reviewed approximately 20 percent of all coded responses, including cases that coders flagged for review because they were either unable to code them or did not know how to code them. Approximately 2 percent of all cases required recoding.

Table II.3. Supplemental Codes for Occupation and Industry Coding

Code	Label	Description
94	Sheltered workshop	The code used if the occupation is in a sheltered workshop and the occupation cannot be coded from verbatim.
95	Refused	The respondent refuses to give his or her occupation or type of business.
97	No occupation or industry reported	No valid occupation or industry is reported in the verbatim response.
98	Don't know	The respondent reports that he or she does not know the occupation or industry.
99	Uncodeable	A code cannot be assigned based on the verbatim response.

This page has been left blank for double-sided copying.

III. SAMPLING WEIGHTS

We determined the final analysis weights for the representative beneficiary sample (RBS) via a four-step process:

1. Calculate the initial probability weights
2. Adjust the weights for two phases of nonresponse (location and cooperation)
3. Trim the weights to reduce the variance
4. Conduct post-stratification

In Section A, we summarize the procedures used to compute and adjust the sampling weights. In Section B, we describe the procedures for computing the weights for the RBS in more detail.

A. Computing and Adjusting the Sampling Weights: A Summary

The sampling weights for any survey are computed from the inverse selection probability that incorporates the stages of sampling in the survey. We selected the RBS in two stages by (1) selecting primary sampling units (PSUs) and (2) selecting the individuals within the PSUs from a current database of beneficiaries.[10] For the prior four rounds of the NBS, we selected PSUs only once (in 2003). By using data from SSA on the counts of eligible beneficiaries in each county, we formed 1,330 PSUs, each of which consisted of one or more counties. The first-stage sampling units in Round 5 of the NBS–General Waves were selected from the same list of PSUs.[11] The PSUs selected in this round will be the first-stage sampling units for all subsequent rounds. We selected 79 of these PSUs, with 2 PSUs—Los Angeles County, California, and Cook County, Illinois—acting as certainty PSUs because of their large size.[12] The Los Angeles PSU received a double allocation because it deserved two selections based on its size relative to other PSUs. The sample of all SSA beneficiaries was selected from among beneficiaries residing in these 79 PSUs. The Los Angeles County and Cook County PSUs had a much larger number of beneficiaries than other counties. Therefore, we partitioned them into a large number of secondary sampling units (SSUs) based on beneficiary zip codes.[13] From these SSUs, we

[10] In two primary sampling units (PSUs), we used an intermediate stage for sampling of secondary sampling units (SSUs). For the sake of simplicity, these SSUs are generally equivalent to PSUs in this description.

[11] Because the geographical distribution of beneficiaries changed little between 2003 and 2014, we kept the same set of 1,330 PSUs that were created for the prior NBS. Although the set of PSUs from which to sample did not change from the prior NBS to the current NBS, we selected a new set of sampled PSUs by using a measure of size for each PSU based on the most current counts of beneficiaries.

[12] Los Angeles County includes the city of Los Angeles; Cook County includes the city of Chicago.

[13] We used the same process for creating and selecting SSUs as we did for the PSUs. Furthermore, we used the same list of SSUs in this round of the current NBS as those created in 2003 for prior to Round 1. But we selected a new set of SSUs for the sample by using a measure of size for each SSU that was based on the most current counts of beneficiaries.

selected four SSUs from the Los Angeles County PSU and two from the Cook County PSU.[14] Beneficiaries were selected from the PSUs or SSUs by using age-defined sampling strata. In total, we selected SSA beneficiaries from 83 locations (77 PSUs and 6 SSUs) from across the 50 states and the District of Columbia. In the remainder of this document, we refer to this set of 83 locations as PSUs.

We sampled beneficiaries in the selected PSUs who were in active pay status as of June 30, 2014.[15] We used four age-based strata in each PSU. In particular, we stratified beneficiaries into the following age groups: (1) 18- to 29-year-olds, (2) 30- to 39-year-olds, (3) 40- to 49-year-olds, and (4) 50-year-olds and older. Because we used a composite size measure to select the PSUs, we could achieve equal probability samples in the age strata and nearly equal workload in each PSU for the RBS.[16]

For the initial beneficiary sample, we selected more individuals than we expected to need in order to account for differential response and eligibility rates in both the PSUs and the sampling strata. We randomly partitioned this augmented sample into subsamples (called "waves") and used some of the waves to form the actual final sample (that is, the sample released for data collection). We released an initial set of waves and then monitored data collection to identify which PSUs and strata required additional sample members. After we released sample members in the initial waves, we were able to limit the number of additional sample members (in subsequently released waves) to those PSUs and strata that required them. Thus, we achieved sample sizes close to our targets while using the smallest number of beneficiaries. Controlling the release of the sample also allowed us to control the balance between data collection costs and response rates. We computed the initial sampling weights based on the inverse of the selection probability for the augmented sample. Given that we released only a subset of the augmented sample, we then adjusted the initial sampling weights for the actual sample size. The release-adjusted weights were post-stratified to population totals that were obtained from SSA.[17] In this report, these release-adjusted sampling weights are referred to as the base weights.

We then needed to adjust the base weights for nonresponse. A commonly used method for computing weight adjustments is to form classes of sample members with similar characteristics and then use the inverse of the class response rate as the adjustment factor in that class. The

[14] It was possible for a beneficiary to reside in one of the selected PSUs (Los Angeles County or Cook County) and not be selected because the beneficiary did not reside in one of the selected SSUs.

[15] We included SSI beneficiaries with selected nonpayment (PSTAT) status codes only if the denial variable (DENCDE) was blank. These are suspension codes that could return to current pay if the beneficiary's application was not in a denial status. During the data collection period, beneficiaries who were found to be deceased, incarcerated, no longer living in the continental United States, or who reported that they had not received benefits in the past five years at the time of the interview, were marked as ineligible. The proportion of cases marked as ineligible during data collection (4.0 percent) was lower than the ineligibility rates obtained in the prior rounds (6.0 percent in Round 4, 6.4 percent in Round 3, 5.6 percent in Round 2, and 5.1 percent in Round 1). The impact on yield rates was negligible.

[16] The composite size measure was computed from the sum of the products of the sampling fraction for a stratum and the estimated count of beneficiaries in that stratum and PSU (Folsom et al. 1987).

[17] The totals were obtained from a frame file provided by SSA that contained basic demographics for all SSI and SSDI beneficiaries.

adjusted weight is the product of the base weight and the adjustment factor. One would form the "weighting classes" to ensure that there would be sufficient counts in each class to make the adjustment more stable (that is, to ensure smaller variance). The natural extension to the weighting class procedure is to perform logistic regression with the weighting class definitions used as covariates, provided that each level of the model covariates has a sufficient number of sample members to ensure a stable adjustment. The inverse of the propensity score is then the adjustment factor. The logistic regression approach also has the ability to include both continuous and categorical variables; standard statistical tests are available to evaluate the selection of variables for the model. For the nonresponse weight adjustments (at both the location and cooperation stages), we used logistic models to estimate the propensity for a sample member to respond. The adjusted weight for each sample case is the product of the base weight and the adjustment factor.

We calculated the adjustment factor in two stages: (1) by estimating a propensity score for locating a sample member and (2) by estimating a propensity score for response among these located sample members. In our experience with the NBS, factors associated with the inability to locate a person tend to differ from factors associated with cooperation. The unlocated person generally does not deliberately avoid or otherwise refuse to cooperate. For instance, that person may have chosen not to list his or her phone number or may frequently move from one address to another, but there is no evidence to suggest that once located he or she would show a specific unwillingness to cooperate with the survey. Located nonrespondents, on the other hand, may deliberately avoid the interviewer or express displeasure or hostility toward surveys in general or toward SSA in particular.

To develop the logistic propensity models for this round, we used as covariates information from the SSA data files as well as geographic information (such as urban or rural region). We obtained much of the geographic information from the Area Health Resource File (AHRF 2014), a file with county-level information on population, health, and economic-related matters for every county in the United States. By using a liberal level of statistical significance (0.3) in forward and backward stepwise logistic regression models, we made an initial attempt to reduce the pool of covariates and interactions. We used a higher significance level because each model's purpose was to improve the estimation of the propensity score, not to identify statistically significant factors related to response. In addition, the information sometimes reflected proxy variables for some underlying variable that was both unknown and unmeasured. We excluded from the pool any covariate or interaction that was clearly unrelated to locating the respondent or to response propensity. Given that the stepwise logistic regression analysis does not fully account for the complex survey design, we developed the final weighted models by using SUDAAN software, which accounts appropriately for the complex sample design.

The next step called for the careful evaluation of a series of models by comparing the following measures of predictive ability and goodness of fit: the R-squared statistic, Akaike's Information Criterion (AIC)[18], the percentage of concordant and discordant pairs, and the

[18] Akaike's Information Criterion is defined as AIC = $-2LogL + 2(k + s)$, where $LogL$ is the log likelihood of the binomial distribution using the parameters from the given model, k is the total number of response levels minus 1, and s is the number of explanatory effects (Akaike 1974). AIC is a relative number and has no meaning on its own. For a given model, smaller values of AIC are preferable to larger values.

Hosmer-Lemeshow Goodness-of-Fit Test. Model-fitting also involved reviewing the statistical significance of the coefficients of the covariates in the model and avoiding any unusually large adjustment factors. In addition, we manipulated the set of variables to avoid data warnings in SUDAAN.[19] We then used the specific covariate values for each located person to estimate the propensity score, from which the adjustment factor was determined by taking the inverse. When computing the adjustment factors, we reviewed their distribution to identify and address any adjustment factors that were outliers (very large or very small relative to other adjustment factors). The location-adjusted weight is the product of the released-adjusted probability weight and the location adjustment. The nonresponse-adjusted weight is the product of the location-adjusted weight and the inverse of the cooperation propensity score, calculated in the same manner as the location propensity score.

Once we made the adjustments, we assessed the distribution of the adjusted weights for unusually high values, which could make the survey estimates less precise. We used the design effect attributed to the variation in the sampling weights as a statistical measure to determine both the necessity and amount of trimming. The design effect attributed to weighting is a measure of the potential loss in precision caused by the variation in the sampling weights relative to a sample of the same size with equal weights. We also wanted to minimize the extent of trimming to avoid the potential for bias in the survey estimates. For the RBS, we checked the design effect attributable to unequal weighting within the age-related sampling strata and determined that no further trimming of the adjusted weights was required. The maximum design effect among all age strata in the RBS was 1.08.

The final step is a series of post-stratification adjustments through which the weights sum to known totals obtained from SSA on various dimensions—specifically, gender, age grouping, program title,[20] and five categories of annual earnings from the Disability Control Files (DCF) of 2013 and 2014.[21] After post-stratification, we checked the survey weights again to determine

[19] SUDAAN data warnings usually included one or more of the following: (1) an indication of a response cell with a zero count; (2) one or more parameters approaching infinity, which may not be readily observable with the parameter estimates themselves; and (3) degrees of freedom for overall contrast that were less than the maximum number of estimable parameters. We tried to avoid all of these warnings, although avoidance of the first two was of highest priority. The warnings usually were caused by a response cell with a count that was too small, which required dropping covariates or collapsing categories in covariates.

[20] Disability payments were made in the form of SSI or SSDI or both.

[21] This was an attempt to address small negative bias in annual earnings, which was observed in past rounds. We arrived at the five earnings categories, which are given in Table III.2, after a lengthy investigation using both (annual) IRS and (monthly) DCF earnings. Using data from the 2014 sampling frame, we calculated the percent with positive IRS earnings in 2014 (considered as "working"), as well as the mean and median IRS 2014 earnings, both overall and among those who were working. We compared these values to several sets of poststratified weights, where the poststratification was based on a variety of earnings categorical variables, each with different cutpoints, some with IRS earnings and some with DCF earnings. We determined that, although the IRS earnings are more accurate than DCF earnings, IRS earnings are only available annually, raising timing issues, and diluting the advantage of accuracy. It was also more difficult to use IRS earnings, since they could only be accessed by staff at SSA. We arrived at the cutpoints given above because these cutpoints resulted in a poststratified weights that resulted in estimated annual earnings that were closest to the IRS values. The 2013 data was used because of a lag in identifying earnings in the 2014 data, which did not have complete information on the amount of earnings that beneficiaries received in that year.

whether more trimming was needed. In this round, trimming was not needed after post-stratification in the RBS.

1. Quality Assurance

To ensure that the methods used to compute the weights at each step were sound, a senior statistician conducted a final quality assurance check of the weights from the representative beneficiary cross-sectional samples. For the sake of objectivity, we chose a statistician who was not directly involved in the project.

B. Details of Calculation of Weights

1. Base Weights

We computed the initial sampling weights by using the inverse of the probability of selection. For the RBS, we selected samples independently in each of four age strata in each PSU. We determined the number of sample members selected in each stratum and PSU for the augmented sample by independently allocating four times the target sample size across the 83 PSUs for each stratum,[22] thereby ensuring the availability of ample reserve sample units in case response or eligibility rates were lower than expected. The augmented sample size for the three younger age strata (18- to 29-year-olds, 30- to 39-year-olds, and 40- to 49-year-olds) was 4,444 sample members (roughly four times the target sample size of 1,111). For beneficiaries age 50 and older, the augmented sample size was 2,667 (again, about four times the target sample size of 667). By using the composite size measure already described, we calculated the initial weights for the full augmented sample of 15,999 sample members by taking the inverse of the global sampling rate (Fj) for each stratum. In Table III.1, we provide the global sampling rates and initial weights, as well as the sizes of the population, augmented sample, and released sample.

Table III.1. Study Population (as of June 30, 2014), Initial Augmented Sample Sizes, and Initial Weights by Sampling Strata in the National Beneficiary Survey

Sampling Strata (ages as of June 30, 2015)	Study Population	Augmented Sample Size	Global Sampling Rate (*Fj*)	Initial Sample Weights	Released Sample
Beneficiaries age 18 to 29	1,415,739	4,444	0.003139	318.57	2,268
Beneficiaries age 30 to 39	1,453,588	4,444	0.003057	327.09	2,126
Beneficiaries age 40 to 49	2,373,419	4,444	0.001872	534.07	2,076
Beneficiaries age 50 to FRA	8,566,947	2,867	0.000335	2,988.1	1,212
Total	**13,809,693**	**15,999**			**7,682**

Source: Study population counts are from SSA administrative CERs and DBADs files. SSA determined the number of complete interviews based upon recommendations from Mathematica.

FRA = full retirement age.

[22] We selected an augmented sample that was four times as large as needed in order to allow for both an adequate supplemental sample in all PSUs and sampling strata within the PSUs and to account for expected variation in the response and eligibility rates across PSUs and sampling strata.

As described previously, we randomly partitioned the full sample into subsamples called "waves" that mirrored the characteristics of the full sample. The waves were formed in each of the four sampling strata in the 83 PSUs (a total of 332 combinations of PSUs and sampling strata). At the start of data collection, we assigned a preliminary sample to the data collection effort and then assigned additional waves as needed, based on experience with eligibility and response rates. Within the 332 combinations of PSUs and sampling strata, we adjusted the initial weights to account for the number of waves released to data collection. The final sample size for the RBS totaled 7,682 beneficiaries, as shown in Table III.1.

2. Response Rates and Nonresponse Adjustments to the Weights

As in virtually all surveys, we had to adjust the sampling weights to compensate for sample members who could not be located or who, once located, refused to respond. First, we fitted weighted logistic regression models where the binary response was whether the sample member could be located. Using variables obtained from SSA databases, we selected, through stepwise regression, a pool of covariates from which to construct a final location model. The pool included both main effects and interactions. From the pool of covariates, we used various measures of goodness of fit and predictive ability to compare candidate models while avoiding large adjustments. We repeated the process for interviewed respondents among the located sample members and fitted another weighted logistic regression model. The two levels in the binary response for this cooperation model were respondent or nonrespondent. For the RBS, a sample member was classified as a cooperating respondent if the sample member or the person responding for the sample member completed the interview (that is, an eligible respondent) or if the sample member was deemed ineligible after sample selection (an ineligible respondent). Ineligible sample members included persons who were never SSA beneficiaries, were in the military at the time of the survey, were incarcerated, had moved outside the United States, or were deceased at the time of the survey. After adjusting the sampling weight by taking the product of the base weight, the location adjustment, and the cooperation adjustment, we checked the distribution of the adjusted weights within each age category and trimmed the weights to remove outliers from the distribution, reallocating the trimmed portion of the outlier weights to other weights within the same age category.

Based on the above procedures, the main factors or attributes affecting our ability to locate and interview a sample member included (1) the sample member's personal characteristics (race, ethnicity, gender, and age); (2) the identity of the payee with respect to the beneficiary; (3) whether the beneficiary and the applicant for benefits lived in the same location; (4) how many phone numbers or addresses were in the SSA files for the beneficiary; (5) the living situation of the beneficiary; and (6) geographic characteristics, including attributes of the county where the beneficiary lived. The following sections detail the steps involved in calculating response rates and adjusting weights for nonresponse.

a. Coding of Survey Dispositions

The Mathematica Survey Management System maintained the status of each sample member during the survey, with a final status code assigned after the completion of all locating and interviewing efforts on a given sample member or at the conclusion of data collection. For the nonresponse adjustments, we classified the final status codes into four categories:

1. Eligible respondents
2. Ineligible respondents (sample members ineligible after sample selection, including deceased sample members, sample members in the military or incarcerated, sample members living outside the United States, and other ineligibles)
3. Located nonrespondents (including active or passive refusals and language barrier situations)
4. Unlocated sample members (sample members who could not be located through either central office tracing procedures or in-field searches)

This classification of the final status code allowed us to measure the location rate among all sample members, the cooperation rate among located sample members, and the overall response rate.

b. Response Rates

The 62.6 percent response rate for the RBS (Table III.2) is the weighted[23] count of sample members who completed an interview or were deemed ineligible divided by the weighted sample count of all sample members.[24] It can be approximated by taking the product of the weighted location rate and the weighted cooperation rate among located sample members.[25]

The weighted location rate is the ratio of the weighted sample count for located sample members to the weighted count of all sample members, which was 88 percent (Table III.2). The weighted cooperation rate (that is, the weighted cooperation rate among located sample members) of 71 percent (Table III.2) is the weighted count of sample members who completed an interview or were deemed ineligible divided by the weighted sample count of all located sample members. Weighted cooperation rates reflect the rate at which completed interviews are obtained from repeated contact efforts among located persons.

[23] This response rate is calculating using the base weight, also referred to as the release-adjusted sampling weight.

[24] The response rate is calculated as the weighted count of sample members who completed an interview or were deemed ineligible divided by the weighted sample count of all sample members: (number of completed interviews + number of partially completed interviews + number of ineligibles)/(number of cases in the sample). The response rate is essentially equivalent to the American Association of Public Opinion Research (AAPOR) standard response rate calculation, assuming that all nonrespondents have unknown eligibility status: RR_{AAPOR} = number of completed interviews/(number of cases in the sample - estimated number of ineligible cases). Ineligible cases are included in the numerator and denominator for two reasons: (1) the cases classified as ineligible are part of the original sampling frame (and hence the study population) and we obtained complete information for fully classifying these cases (that is, their responses to the eligibility questions in the questionnaire are complete) such that we may classify them as respondents; and (2) incorporation of the ineligibles into the numerator and denominator of the response rate is essentially equivalent to the definition of a more conventional response rate, assuming that all nonrespondents have unknown eligibility status.

[25] This product is not exactly equal to the weighted response rate, since the location rate is calculated using the base weight, and the cooperation rate among located cases is calculated using the location-adjusted base weight.

Table III.2. Weighted Location, Cooperation, and Response Rates for Representative Beneficiary Sample, by Selected Characteristics

	Sample	Located Sample		Response Among Located Sample		Overall Respondents
	Count	Count	Location Rate	Count	Cooperation Rate	Response Rate
All	**7,682**	**6,446**	**87.9**	**4,359**	**71.0**	**62.6**
SSI Only, SSDI Only, or Both SSI and SSDI						
SSI only	3,196	2,603	85.6	1,749	69.1	59.2
SSDI only	3,034	2,611	89.3	1,748	71.2	63.8
Both SSI and SSDI	1,452	1,232	87.4	862	73.8	64.6
Constructed Disability Status						
Deaf	83	68	76.9	44	49.3	37.6
Cognitive disability	1,542	1,297	86.5	867	67.1	58.1
Mental illness	2,896	2,363	85.3	1,574	70.4	60.1
Physical disability	2,987	2,584	89.8	1,798	72.8	65.5
Unknown	174	134	86.6	76	57.6	50.0
Beneficiary's Age (four categories)						
18 to 29	2,268	1,820	81.6	1,240	69.0	56.5
30 to 39	2,126	1,762	84.0	1,168	67.8	57.1
40 to 49	2,076	1,779	86.3	1,186	68.2	59.0
50 and older	1,212	1,085	90.1	765	72.6	65.5
Sex						
Male	4,083	3,395	86.7	2247	69.5	60.4
Female	3,599	3,051	89.2	2112	72.5	64.8
Ethnicity (Hispanic or not)						
Hispanic	380	310	88.2	213	73.3	64.7
Non-Hispanic	5,904	5,003	88.2	4,146	70.9	62.6
Race						
White	3,906	3,320	88.5	2,225	70.7	62.8
Black	1,645	1,376	87.3	949	71.5	62.5
Hispanic	380	310	88.2	213	73.3	64.7
Unknown	1,649	1,349	86.2	910	70.9	61.3
Asian American, Pacific Island American,	77	70	92.4	46	68.1	62.7
American Indian, or Alaska Native	25	21	92.6	16	71.6	67.6
Living Situation						
Living alone	4,057	3,347	86.6	2,278	70.7	61.3
Living with others	330	286	86.9	204	75.8	66.3
Living with parents	125	93	78.6	52	59.9	48.6
In institution or unknown	52	48	93.0	30	58.6	55.1
Did the Applicant for Benefits Live in the Same ZIP Code as the Beneficiary?						
No	852	669	80.5	428	63.7	51.1
Yes	5,095	4,315	88.1	2,990	72.9	64.3
No information	1,735	1,462	89.4	941	69.2	62.0

TABLE III.2 (continued)

	Sample	Located Sample		Response Among Located Sample		Overall Respondents
	Count	Count	Location Rate	Count	Cooperation Rate	Response Rate
Identity of the Payee with Respect to the Beneficiary						
Beneficiary received payments directly	330	270	83.6	187	71.1	59.2
Payee is a family member	2,319	1,951	86.9	1,334	70.2	61.0
Payee is an institution	365	303	84.6	180	63.7	54.3
Other	4,668	3,922	88.5	2,658	71.5	63.4
Count of Phone Numbers in File						
Only one phone number in file	660	580	89.7	396	66.2	59.4
Two phone numbers in file	1,142	975	91.0	658	72.1	65.9
Three phone numbers in file	1,516	1,318	90.3	911	71.8	64.8
Four phone numbers in file	1,497	1,271	89.4	881	73.1	65.4
Five phone numbers in file	1,185	971	86.2	667	72.4	62.4
Six or more phone numbers on file	1,674	1,326	82.4	841	67.6	55.8
Count of Addresses in File						
One address in file	780	719	94.9	508	73.9	70.2
Two addresses in file	1,411	1,243	92.0	872	71.3	65.5
Three addresses in file	1,596	1,355	89.1	935	71.9	64.3
Four addresses in file	1,510	1,253	86.4	824	69.5	60.1
Five or more addresses in file	2,382	1,876	83.5	1,220	70.3	58.9
Census Region						
Midwest	1,581	1,389	91.8	966	75.8	69.7
Northeast	1,490	1,258	88.6	814	67.2	59.6
South	3,127	2,583	86.9	1,820	72.2	62.8
West	1,484	1,216	85.3	759	66.3	56.7
Census Division						
East North Central	1,082	950	92.1	657	76.6	70.7
East South Central	719	597	88.6	436	72.1	63.9
Middle Atlantic	1,091	932	89.4	596	66.2	59.3
Mountain	454	366	87.0	258	75.4	65.9
New England	399	326	86.2	218	70.1	60.6
Pacific	1,030	850	84.5	501	61.7	52.3
South Atlantic	1,479	1,222	87.5	840	72.3	63.4
West North Central	499	439	91.0	309	73.6	67.2
West South Central	929	764	84.5	544	72.2	61.1

TABLE III.2 (continued)

	Sample	Located Sample		Response Among Located Sample		Overall Respondents
	Count	Count	Location Rate	Count	Cooperation Rate	Response Rate
Metropolitan Status of County						
Metropolitan areas with population of 1 million or more	3,621	3,070	87.7	1,963	66.1	58.1
Metropolitan areas with population of 250,000 to 999,999	2,048	1,700	87.1	1,173	74.3	64.9
Metropolitan areas with population of fewer than 250,000	915	763	89.4	549	72.9	65.2
Nonmetropolitan areas adjacent to large metropolitan areas	252	220	88.9	162	79.5	71.1
Nonmetropolitan areas adjacent to medium or small metropolitan areas	604	490	88.0	362	77.2	68.0
Nonmetropolitan areas not adjacent to metropolitan areas	242	203	90.4	150	76.7	69.4
County with Low Education						
Yes	938	770	85.2	490	68.7	58.6
No	6,744	5,676	88.3	3,869	71.3	63.1
County with Housing Stress						
Yes	3,094	2,563	85.4	1,616	65.8	66.3
No	4,588	3,883	89.4	2,743	73.9	64.5
Population Loss County						
Yes	395	335	89.3	221	72.1	64.5
No	7,287	6,111	87.9	4,138	70.9	62.5
Retirement Destination County						
Yes	1,139	963	88.8	665	72.4	64.5
No	6,543	5,483	87.8	3,694	70.7	62.2
Service-Dependent Economy County						
Yes	3,207	2,683	86.8	1,742	66.5	57.7
No	4,475	3,763	88.7	2,617	73.9	65.8
Nonspecialized-Dependent Economy County						
Yes	2,013	1,707	89.7	1,176	74.2	66.8
No	5,669	4,739	87.3	3,183	69.8	61.1
Government-Dependent Economy County						
Yes	865	718	87.0	481	69.0	60.2
No	6,817	5,728	88.1	3,878	71.2	62.9

TABLE III.2 (continued)

	Sample	Located Sample		Response Among Located Sample		Overall Respondents
	Count	Count	Location Rate	Count	Cooperation Rate	Response Rate
County Racial/Ethnic Profile						
County with at least 90% non-Hispanic white population	758	657	91.6	453	74.7	68.6
County with plurality or majority Hispanic population	685	569	85.2	358	66.3	56.4
County with majority but fewer than 90% non-Hispanic white population	3,468	2,920	88.4	2,002	72.9	64.6
County with a racially/ethnically mixed population, no majority group	2,561	2,134	87.2	1,438	68.5	59.8
County with plurality or majority non-Hispanic black population	210	166	82.6	108	67.6	55.9
DCF Earnings Category[a]						
Beneficiary with monthly DCF earnings above SGA[b] for three consecutive months in 2013 or 2014	101	85	83.1	49	59.8	50.6
Beneficiary with annual DCF earnings above $7,000 in 2013 or 2014	185	155	90.8	96	67.0	61.0
Beneficiary with annual DCF earnings above $2,000 in 2013 or 2014	289	248	89.1	158	70.6	63.3
Beneficiary with annual DCF earnings above $0 in 2013 or 2014	342	297	90.5	195	71.9	65.6
Beneficiary with no annual DCF earnings in 2013 or 2014	6,765	5,661	87.8	3,861	71.1	62.4

Source: NBS–General Waves Round 5.

[a]The DCF earnings categories are subdivided sequentially. In other words, the second category excludes those who were in the first category; the third excludes those that are in the first or second category, and so on.

[b]Non-blind substantial gainful activity, or $1,070 in 2014 and $1,040 in 2013.

We use the weighted rates because (1) the sampling rates (therefore, the sampling weights) vary substantially across the sampling strata (as seen in Table III.1) and (2) the weighted rates better reflect the potential for nonresponse bias. The weighted rates represent the percentage of the full survey population for which we were able to obtain information sufficient for use in the data analysis or in determining ineligibility for the analysis.

c. Factors Related to Location and Response

In addition to overall response rate information, Table III.2 provides information for factors that were considered for use in the location and cooperation models. The table displays the unweighted counts of all sample members, counts of located sample members, and counts of sample members who completed an interview or who were deemed ineligible. We also include in the table the weighted location rate, the weighted cooperation rate among located sample members, and the weighted overall response rate for these factors, which helped inform the decision about the final set of variables to be used in the nonresponse adjustment models.

d. Propensity Models for Weight Adjustments

Using the main effects already described as well as selected interactions, we developed response propensity models to determine the nonresponse adjustments. To identify candidate interactions from the main effects for the modeling, we first ran a chi-squared automatic interaction detector (CHAID) analysis in SPSS to find possible significant interactions.[26] The CHAID procedure iteratively segments a data set into mutually exclusive subgroups that share similar characteristics based on their effects on nominal or ordinal dependent variables. It automatically checks all variables in the data set and creates a hierarchy showing all statistically significant subgroups. The algorithm identifies splits in the population, which are as different as possible based on a chi-squared statistic. The forward stepwise procedure finds the most diverse subgroupings and then splits each subgroup further into more diverse sub-subgroups. Sample size limitations are set to avoid cells with small counts. The procedure stops when splits are no longer significant; that is, a group is homogeneous with respect to variables not yet used or the cells contain too few cases. The CHAID procedure produces a tree that identifies the set of variables and interactions among the variables that are associated with the ability to locate a sample member (and a located sample member's propensity either to respond to or to be deemed ineligible for the NBS). We first ran CHAID with all covariates and then reran it a few times with the top variable in the tree removed to ensure the retention of all potentially important interactions for additional consideration. We further reduced the resulting pool of covariates by evaluating tabulations of all the main effects and the interactions identified by CHAID. At a particular level of a given covariate or interaction, if all respondents were either located or unlocated (for the location models), complete or not complete (for the cooperation models), or the total number of sample members at that level was fewer than 20, the levels were collapsed if

[26] CHAID is normally attributed to Kass (1980) and Biggs et al. (1991). Its application in SPSS is described in Magidson (1993).

collapsing was possible. If collapsing was not possible, then we excluded the covariate or interaction from the pool.[27]

To further refine the candidate variables and interaction terms, we processed all of the resulting candidate main effects and the interactions identified by CHAID using forward and backward stepwise regression (using the STEPWISE option of the SAS LOGISTIC procedure with weights normalized to the sample size).[28] After identifying a smaller pool of main effects and interactions for potential inclusion in the final model, we carefully evaluated a set of models to determine the final model. Given that the SAS logistic regression procedure does not incorporate the sampling design, we relied on the logistic regression procedure in SUDAAN to make the final selection of covariates.

For selecting variables or interactions in the stepwise procedures, we included variables or interactions with a statistical significance level (alpha level) of 0.30 or lower (instead of the commonly used 0.05).[29] Once we determined the candidate list of main effects and interactions, we used a thorough model-fitting process to determine a parsimonious model with few very small propensities. (In Section A of this chapter, we described the model selection criteria.) Once we decided which interactions to include in each final model, the main effects corresponding to each interaction were also included in the final model, regardless of the significance level of those main effects. For example, suppose the age by gender interaction was significant in the location model. In that case, the significance levels for the age and gender main effects were not important, because the nature of the relationship between location, age, and gender is contained in the interaction. In Table III.3, we summarize the variables used in the model as main effects and interactions for locating a sample member. In Table III.4, we summarize the variables used in the model for cooperation among located sample members.

[27] Deafness historically has been shown to be an important indicator both of locating a sample member and determining whether the sample member completed the interview. For that reason, deafness remained in the covariate pool even though the number of deaf cases was sometimes as few as 18.

[28] SUDAAN offers no automated stepwise procedures; the stepwise procedures described here were performed by using SAS.

[29] As stated, we used a higher significance level because the model's purpose was to improve the estimation of the propensity score rather than to identify statistically significant factors related to response. In addition, the information sometimes reflected proxy variables for some underlying variable that was both unknown and unmeasured.

Table III.3. Location Logistic Propensity Model: Representative Beneficiary Sample

Factors in Location Model
Main Effects
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
GENDER (MALE OR FEMALE)
AGECAT (AGE CATEGORY)
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
CNTYNONSP (NONSPECIALIZED-DEPENDENT ECONOMY COUNTY)
Two-Factor Interactions
PHONE*CNTYRACE

Table III.4. Cooperation Logistic Propensity Model: Representative Beneficiary Sample

Factors in Cooperation Model
Main Effects
AGECAT (AGE CATEGORY)
RACE
DISABILITY
METRO (METROPOLITAN STATUS OF COUNTY)
GENDER (SEX)
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
CNTYSVC (SERVICE-DEPENDENT ECONOMY COUNTY)
CNTYGOV (GOVERNMENT-DEPENDENT ECONOMY COUNTY)
CNTYLOWEDUC (LOW-EDUCATION COUNTY)
CNTYPERSPOV (COUNTY WITH PERSISTENT HIGH LEVELS OF POVERTY)
CNTYHSTRESS (COUNTY WITH HIGH LEVELS OF HOUSING THAT WAS OF POOR QUALITY, CROWDED, AND/OR EXPENSIVE RELATIVE TO INCOME LEVELS)
Two-Factor Interactions
PDZIPSAME*CNTYHSTRESS
PDZIPSAME*CNTYPERSPOV
PDZIPSAME*CNTYSVC
CNTYHSTRESS*METRO
CNTYHSTRESS*PHONE
CNTYSVC*PHONE

The R-squared is 0.029 (0.055 when rescaled to have a maximum of 1) for the location model and 0.042 (0.0607 when rescaled) for the cooperation model.[30] These values are similar to those observed for other response propensity modeling efforts that use logistic regression with design-based sampling weights. For the location model, 61 percent of pairs are concordant, 36.9 percent of pairs are discordant,[31] and the p-value for the chi-square statistic from the Hosmer-Lemeshow (H-L) Goodness-of-Fit Test is 0.128.[32] These values indicate a reasonably good fit of the model to the data. The location adjustment from the model, calculated as the inverse of the location propensity score, ranged from 1.00 to 1.72. For the cooperation model, 56.9 percent of pairs are concordant and 42.4 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.678 for the model. The cooperation adjustment from the model, which is calculated as the inverse of the cooperation propensity score, ranged from 1.04 to 4.09. The overall nonresponse adjustment (the product of the location adjustment and the cooperation adjustment) ranged from 1.07 to 4.87.[33]

Among the variables used in the location and cooperation models shown in Tables III.3 and III.4, the number of levels used in the models is often fewer than the number of levels in Table III.2; the levels collapsed for the models are described following the tables. The factors used in the location model included the following:

- **MOVE.** Count of addresses in SSA files. There are five levels: (1) one address in file, (2) two addresses in file, (3) three addresses in file, (4) four addresses in file, (5) five or more addresses in file or no information.
- **PHONE.** Count of phone numbers in SSA files. There are three levels: (1) one to three phone numbers in file, (2) four or more phone numbers in file, or (3) no information.
- **GENDER.** Beneficiary's sex. There are two levels: (1) male and (2) female.
- **PDZIPSAME.** Whether the beneficiary and the applicant for benefits lived in the same ZIP code. There are two levels: (1) beneficiary and applicant lived in different ZIP codes and (2) beneficiary and applicant lived in same ZIP codes or the information is unknown.
- **AGECAT.** Beneficiary's age category. There are four levels: (1) age 18 to 29, (2) age 30 to 39, (3) age 40 to 49, (4) age 50 or older.

[30] The Generalized Coefficient of Determination (Cox and Snell 1989) is a measure of the adequacy of the model, in which higher numbers indicate a greater difference between the likelihood of the model in question and the null model. The Max Rescaled R-Square scales this value to have a maximum of 1.

[31] A pair of observations is concordant if a responding subject has a higher predicted value than a nonresponding subject, discordant if not, and tied if both members of the pair are respondents, nonrespondents, or have the same predicted values. It is desirable to have as many concordant pairs and as few discordant pairs as possible (Agresti 1996).

[32] The Hosmer-Lemeshow Goodness-of-Fit Test is a test for goodness of fit of logistic regression models. Unlike the Pearson and deviance goodness-of-fit tests, it may be used to test goodness of fit even when some covariates are continuous (Hosmer and Lemeshow 1989).

[33] Recognizing that the Akaike's Information Criterion is a relative number and has no meaning on its own, we do not provide values for it here.

- **CNTYNONSP.** Nonspecialized-dependent county. There are two levels: (1) the county's economy is not dependent upon farming, mining, manufacturing, government, or services; and (2) the county's economy is dependent upon farming, mining, manufacturing, government, or services, or there is no information.

The model also included various interactions among these variables for locating sample members. In Table III.3, we provide the main effects using the variable names listed above as well as interactions. In Appendix D, we provide an expanded form of Table III.3 showing the levels of interactions shown in Table III.3 along with parameter estimates and their standard errors. The factors used in the cooperation model included the following:

- **AGECAT.** Beneficiary's age category. There are three levels: (1) age 30 to 39, (4) age 40 to 49, (3) age 18 to 29 or age 50 or older.
- **RACE.** Race of beneficiary. There are two levels: (1) non–Hispanic white and (2) not non–Hispanic white or not known to be non–Hispanic white.
- **DISABILITY.** Beneficiary's disability. There are four levels: (1) cognitive disability, (2) mental illness, (3) physical disability (not deafness), (4) deafness or disability unknown.
- **METRO.** Metropolitan status of beneficiary's county of residence. There are six levels: (1) beneficiary lived in metropolitan area with population of 1 million or more; (2) beneficiary lived in metropolitan area with population between 250,000 and 1 million; (3) beneficiary lived in metropolitan area with population fewer than 250,000; (4) beneficiary lived in nonmetropolitan area adjacent to a metropolitan area of 1 million or more; (5) beneficiary lived in nonmetropolitan area adjacent to a metropolitan area of fewer than 1 million; and (6) beneficiary lived in nonmetropolitan area not adjacent to metropolitan area.
- **GENDER.** Beneficiary's sex. There are two levels: (1) male and (2) female.
- **PDZIPSAME.** Whether the beneficiary and the applicant for benefits lived in the same zip code. There are three levels: (1) beneficiary and applicant lived in same zip code, (2) beneficiary and applicant lived in different zip codes, and (3) information unknown.
- **PHONE.** Count of phone numbers in SSA files. There are three levels: (1) one phone number in file, (2) between two and six phone numbers in file, and (3) more than six phone numbers in file or there is no information.
- **CNTYSVC.** County with service-dependent economy. There are two levels: (1) a county with 45 percent or more of average annual labor and proprietors' earnings derived from services (Standard Industrial Classification categories of retail trade; finance, insurance, and real estate; and services) during 1998–2000; and (2) a county without this attribute.
- **CNTYGOV.** County with government-dependent economy. There are two levels: (1) a county where 15 percent or more of average annual labor and proprietors' earnings were derived from federal and state government during 1998–2000, and (2) a county without this attribute.

- **CNTYLOWEDUC.** County with low education. There are two levels: (1) a county where 25 percent or more of residents age 25 through 64 had neither a high school diploma nor a general equivalency diploma (GED) in 2000 and (2) a county without this attribute.
- **CNTYPERSPOV.** County with persistent high levels of poverty. There are two levels: (1) a county where 20 percent or more of residents were poor as measured by each of the last four censuses (1970, 1980, 1990 and 2000); and (2) a county without this attribute.
- **CNTYHSTRESS.** County with high levels of housing that was of poor quality, crowded, or expensive relative to income levels. There are two levels: (1) a county where 30 percent or more of households had one or more adverse housing conditions in 2000 (lacked complete plumbing, lacked complete kitchen, paid 30 percent or more of income for owner costs or rent, or had more than 1 person per room); and (2) a county without this attribute.

Once again, we included various interactions among these variables in the model for the cooperation of sample members. In Table III.4, we provide the main effects using the variable names as well as interactions. In Appendix D, we provide an expanded form of Table III.4, with the levels of the interactions shown in Table III.4 along with parameter estimates and their standard errors.

After we applied adjustments to the sampling weights, we reviewed the distribution of weights to determine the need for further trimming of the weights. We concluded that no additional trimming was needed and that the maximum design effect attributable to unequal weighting was 1.08, which was observed with the youngest age-group stratum.

3. Post-Stratification

Post-stratification is the procedure that aligns the weighted sums of the response-adjusted weights to known totals external to the survey. The process offers face validity for reporting population counts and has some statistical benefits. For the RBS, we post-stratified to the marginal population totals for four variables obtained from SSA. In particular, the totals were the total number of SSI and SSDI beneficiaries by age (four categories); gender; recipient status (SSI only, SSDI only, and both); and DCF earnings (four categories derived from DCF earnings in 2013 and 2014). We conducted no trimming after post-stratification.

This page has been left blank for double-sided copying.

IV. IMPUTATIONS

The data collection instruments for the NBS—General Waves were administered with computer-assisted interviewing technology. The technology allows the use of automated routing to move the respondent to the applicable questions and performs checks of the entered data for consistency and reasonableness. In addition, it does not permit a question to be left blank; therefore, the interviewer may not proceed until an appropriate response has been entered. ("Don't know" and "refused" are included as response options and used as necessary). These processes substantially reduce the extent of item nonresponse for a complex survey, although some item nonresponse will persist—for example, when a question was mistakenly not asked and when "don't know" or "refused" were recorded as responses.

For the NBS—General Waves, we used primarily two methods of imputation to compensate for item nonresponse: (1) deductive (or logical) imputation and (2) unweighted hot-deck imputation. However, for some variables, the data were insufficient to use either method; thus, we needed to employ other methods, such as random draws of imputed values from distributions given by the nonmissing data. Selection of the methods was based on (1) the type of variable (dichotomous, categorical, or continuous); (2) the amount of missing data; and (3) the availability of data for the imputations. For some variables, imputations were processed using a combination of methods.

Deductive imputation is based on a review of the data related to the imputed variable. It assigns a value that may be deduced from other data or for which there is a high degree of certainty that the value is correct.

Hot-deck imputation involves the classification of sample members into mutually exclusive and exhaustive imputation classes (or imputation cells) of respondents who are assumed to be similar relative to the key population variables (such as age, disability status, and SSI recipient status). For each sample member with a missing value (a recipient), a sample member with complete data (a donor) is chosen within the same imputation class to provide a value. Ideally, the imputation class should contain sufficient sample members to avoid the selection of a single donor for several sample members with missing data.

The hot-deck procedure is computationally efficient. A simulation study by the National Center for Education Statistics (U.S. Department of Education 2001) showed that a hot-deck procedure fared well in comparison to more sophisticated imputation procedures, including multiple imputation, Bayesian bootstrap imputation, and ratio imputation. The U.S. Department of Education (USDE) study evaluated imputation methods in terms of bias of the mean, median, and quartile, as well as variance estimates, coverage probability, confidence interval width, and average imputation error.

Although the variance of estimates was a key item used to evaluate methods by the USDE study, we made no attempt in this study to estimate the component of variance attributable to imputation, even though such a component is always positive. Users should be aware that variance estimates that use imputed data will be underestimates, with the amount of bias in the variance estimate directly related to the amount of "missingness" in the variable of interest. For

most of the variables requiring imputation, the extent of missingness was low; thus, the component of variance would be very small in most cases.

For the NBS—General Waves, the hot-deck imputation procedure used an unweighted selection process to select a donor, with selections made within imputation classes that were defined by key related variables for each application. In addition to the variables defining the imputation classes, we included a sorting variable that sorted the recipient and all donors within the imputation class together by levels of the variable. Using the sorted data within the imputation class, we randomly selected as the donor with equal probability a case immediately preceding or following a sample member with missing data. Therefore, the hot-deck procedure was unweighted and sequential, with a random component. We allowed with-replacement selection of a donor for each recipient. In other words, a sample member could have been a donor for more than one recipient. Given that the extent of missing values was very low for most variables, we used only a few donors more than once.[34]

Where appropriate, we made imputed values consistent with pre-existing nonmissing variables by excluding donors with potentially inconsistent imputed values. After processing each imputation, we used a variety of quality control procedures to evaluate the imputed values. If the initial imputed value was beyond an acceptable range or inconsistent with other data for that case, we repeated the imputation until the imputed value was in range and consistent with other reported data.

The factors used to form the cells for each imputed variable needed to be appropriate for the population, the data collected, and the purpose of the NBS—General Waves. In addition, the imputation classes needed to possess a sufficient count of donors for each sample member with missing data. We used a variety of methods to form the imputation classes: bivariate cross-tabulations, stepwise regressions, and multivariate procedures such as CHAID.[35] To develop the imputation classes, we used information from both the interview and SSA administrative data files. The classing and sorting variables were closely related to the variable to be imputed (the response variable). The sorting variables were either less closely related to the response variable than were the classing variables or were forms of the classing variables with finer levels. As an example of the latter situation, we sometimes used four age categories as imputation classes: (1) 18- to 29-year-olds, (2) 30- to 39-year-olds, (3) 40- to 49-year-olds, and (4) those who were 50 years old or older. We could then use the actual age as a sorting variable to ensure that donors and recipients were as close together in age as possible.

In the case of missing values in the variables used to define imputation classes, we applied two strategies: (1) matching recipients to donors who were also missing the value for the covariate or (2) employing separate hot decks, depending upon the availability of the variables defining the imputation classes. In the first instance, we treated the level defined as the missing value as a separate level. In other words, if a recipient was missing a value for a variable defining

[34] Household income, which was used to determine the federal poverty threshold indicator, was the exception. About 17 percent of respondents gave no household income information at all and about 18 percent gave only general categories of income. Detailed levels of missingness are given for all imputed variables later in this chapter.

[35] Chi-Squared Automatic Interaction Detection software is attributed to Kass (1980) and Biggs et al. (1991). Its application in SPSS is described in Magidson (1993).

an imputation class, the donor also was missing the value for that variable. We used the first strategy if a large number of donors and recipients were missing the covariate in question. In the second instance, we used a variable for a given recipient to define the imputation class for that recipient only if there was no missing value for that variable. The variables used to define an imputation class for each recipient depended upon what values were not missing among those variables.

The hot-deck software automatically identified situations in which the imputation class contained only recipients and no donors. In such cases, we collapsed imputation classes and once again performed the imputation with the collapsed classes. The strategy for collapsing classes required a ranking of the variables used to define the imputation class with regard to each variable's relationship to the variable requiring imputation. If several covariates aided in imputing a given variable, the covariates less closely related to the variable requiring imputation were more likely than the important covariates in the imputation to have levels that we had to collapse. In addition, variables with a large number of levels also were more likely to have levels that we had to collapse. In general, if more than a very small number of imputation classes required collapsing, we dropped one or more variables from the definition of the imputation class and reran the imputation procedure.

Some variables were constructed from two or more variables. For some of the constructed variables, it was more efficient to impute the component variables and then impose the recoding of the constructed variable on these imputed values, rather than imputing the constructed variable directly. In the tables that follow in this chapter, we do not show the component variables because they were not included in the final data set.

For some imputed variables in the data set, the number of missing responses does not match the number of imputed responses. Often, the variables correspond to questions that follow a filter question. For example, Item I29 asks if the respondent has serious difficulty walking or climbing stairs. If the response is "yes," the follow-up question (Item I30) asks if the respondent is able to walk without assistance at all. To be asked the follow-up question, the respondent must have answered "yes" to the screener question. If the respondent answered "no," the follow-up question was coded a legitimate missing (.), which was not imputed. However, if the respondent refused to answer the screener question, the follow-up question was also coded a legitimate missing. If the screener variable was then imputed to be "yes," the response to the follow-up question was imputed, causing the count of the actual number of imputed responses to be greater than the number of missing or invalid responses.

A. NBS Imputations of Specific Variables

In the tables below, we present information on how imputation was applied to selected variables in the NBS—General Waves, including the imputed variable names, a brief description of each variable, the methods of imputation, total number of missing responses, number of respondents eligible for the question, and percentage of imputed responses. We recorded this information in the final file with an imputation flag, identified by the suffix "iflag," which has the following levels: (.) legitimate missing, (0) self-reported data, (1) logical imputation, (2) administrative data, (3) hot-deck imputed, (4) imputation using the distribution of a variable related to the variable being imputed, (5) imputation based on specialized procedures specific to

Section K, and (6) constructed from other variables with imputed values. The distinction between "logical imputation" and "constructed from other variables with imputed values" is somewhat opaque. In general, if we made a logical assignment for variables corresponding directly to items from the questionnaire, we set the flag to 1. For variables constructed from these variables (constructed variables are prefixed with a "C_"), we set the flag to 6. In this instance, we imputed one or more of the component variables in the constructed variable. All variables that include imputed values are identified with the suffix "_i."

Below, we summarize the imputations that we conducted and provide details for some of the imputation types for each section of the questionnaire.

1. Section L: Race and Ethnicity

Two items in the questionnaire, item L1 and item L2, gathered information on respondents' race and ethnicity. The imputations associated with these variables are summarized in Table IV.1. In particular, L1_i corresponds to the question asking whether the respondent is Hispanic or not; C_Race_i corresponds to the question asking about the respondent's race.

Table IV.1. Race and Ethnicity Imputations

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
L1_i	Hispanic/Latino ethnic origins	2 imputations from SSA's administrative data, 75 imputations from hot deck	77	4,062	1.90
C_Race_i	Race	78 imputations from SSA's administrative data, 168 imputations from hot deck	246	4,062	6.06

Source: NBS–General Waves Round 5.

Note: The "number missing" is a count of item nonrespondents, and the "number eligible" includes both item respondents and item nonrespondents. The "percentage imputed" is the "number missing" divided by the "number eligible", and is unweighted.

In the above table, respondents who did not indicate in the questionnaire whether they were Hispanic were classified as such if the SSA administrative data so indicated. We also looked at the name of the respondent and compared it to a list of Hispanic names provided by the North American Association of Central Cancer Registries (NAACCR 2003), though in this round no respondents were classified as Hispanic using this method who hadn't already been classified as such using questionnaire or administrative data. For respondents who still had missing data, we imputed the Hispanic indicator by using a hot deck with imputation classes defined by the zip code of each sample member, with race as a sorting variable. Not surprisingly, the imputation classes based on zip code commonly required collapsing to ensure that an imputation class had a sufficient number of donors for the recipients in that class. An automated process in SAS performed the needed check. However, to ensure that the zip code imputation classes being collapsed were as similar as possible, we manipulated the software so that the county of the donor zip code and county of the recipient zip code had a similar racial and ethnic composition according to data from the Area Health Resource File (2014–2015), a file with demographic, health, and economic-related data for every county in the United States.

Respondents could choose from five race categories—(1) white, (2) black/African American, (3) Asian, (4) native Hawaiian or other Pacific Islander, and (5) Alaska native or American Indian—and could select more than one of the categories to identify themselves (as prescribed by the Office of Management and Budget). The final race variable on which imputation was applied included six categories, with a separate category for respondents who reported multiple races. Although the SSA administrative data did not have a category for multiple races, respondents with race information in the SSA files were categorized according to four of the five categories above (native Hawaiian or other Pacific Islanders were included with respondents who reported being Asian). Respondents who did not answer the race question but did have race information in the SSA files were categorized into one of the four categories. This would have resulted in the misclassification of respondents—with SSA administrative data—who did not answer the race question in the survey but who would have identified themselves as multiple race or native Hawaiian or other Pacific Islander. However, we assumed that the number of such respondents would be small and that their misclassification would not be a major problem. As with the Hispanic indicator, for respondents who still had missing data, we imputed race by using a hot deck with imputation classes that were defined by the zip code of each sample member, with ethnicity (Hispanic or not) as a sorting variable.

2. Section B: Disability Status Variables and Work Indicator

Questions about disability status and work were limited to individuals who indicated in Item B1 that they have a "physical or mental condition limiting the kind or amount of work or other daily activities that [they] can do." If the respondent did not answer Item B1, then we imputed Item B1. In this round, there were 11 such cases, 6 of which were imputed as a "1."

In Table IV.2, we describe five imputed variables that pertain to the sample member's disability status and an indicator of whether the respondent was currently working. The imputed variables include three that collapse and recode primary diagnosis codes from the ICD-9 in three ways: (1) C_MainConBodyGroup_i, which corresponds to the collapsing in Table II.2; (2) C_MainConDiagGrp_i; and (3) C_MainConColDiagGrp_i. Additional variables for disability status include age when the disability was first diagnosed (C_DisAge_i) and an indicator of childhood or adult onset of the disability (C_AdultChildOnset_i), variables which were assigned to all survey respondents (not just those with a value of B1 = 1). We also imputed a fourth variable with collapsed primary diagnosis codes, with levels further collapsed from C_MainConDiagGrp_i. Table IV.2 does not include this variable (C_MainConImput_i) because it was not released to the final file but was used in subsequent imputations as a classing variable. Table IV.2 also omits the imputed version of Item B1 (B1_i), as this variable is a supporting variable that was also not released to the final file. All missing values for C_AdultChildOnset_i were "logically assigned" by using the imputed values from C_DisAge_i, the variable for age of onset. In addition, Section B contains a question asking whether the respondent was currently working (Item B24_i), which is a gate question for all of Section C's variables for work status.

Table IV.2. Disability Status Imputations

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
C_MainConDiagGrp_i	Primary diagnosis group	41 hot deck[a]	41	3,583	1.14
C_MainConColDiagGrp_i	Main condition diagnosis group collapsed	41 constructed from imputed variables[a]	41	3,583	1.14
C_MainConBodyGroup_i	Main condition body group	2 hot deck, 39 constructed from imputed variables[a]	41	3,583	1.14
C_DisAge_i	Age at onset of disability	165 hot deck	165	4,062	4.06
C_AdultChildOnset_i	Adult/child onset of disability	10 constructed from imputed variables	10	4,062	0.25
B24_i	Currently working	3 hot deck	3	4,062	0.07

Source: NBS–General Waves Round 5.

Note: The "number missing" and "number eligible" counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The "number missing" is a count of item nonrespondents, and the "number eligible" includes both item respondents and item nonrespondents. The "percentage imputed" is the "number missing" divided by the "number eligible", and is unweighted.

[a]Imputations for diagnosis group variables excluded five cases coded as "don't know" or "refused" in Item B1, which were imputed in Item B1_i as not having a condition that limited the kind or amount of work or other daily activity that the respondent could do.

To define imputation classes, all of the variables in Section B used an indicator to specify whether the onset of the disability occurred in childhood or adulthood and to specify age and gender. We also used one of the collapsed condition code variables, C_MainConImput_i, as a classing variable for disability age and the work indicator. We used additional classing variables specific to the variable being imputed.

3. Section C: Current Jobs Variables

Several survey questions asked respondents about current employment. Section C asked such questions only of respondents who indicated in Item B24 that they were currently working. If the respondent did not answer Item B24, then we imputed Item B24. In this round, there were 2 such cases, both of which were imputed as a "not working." As identified in Table IV.3, the questions asked about the following:

- Salary (C_MainCurJobHrPay_i, C_MainCurJobMnthPay_i, and C_TotCurJobMnthPay_i)
- Usual hours worked at the job or jobs (C8_1_i, C_TotCurWkHrs_i, and C_TotCurHrMnth_i)
- Number of places the respondent was employed (C1_i)
- Job description for the place of main employment (C2_1_1d_i)

We imputed values for other variables by using the distribution of a variable related to the variable at hand. For example, if the take-home monthly pay of the respondent's current main job was not missing but the gross monthly pay (C_MainCurJobMnthPay_i) for the job was missing, we used the relationship between gross monthly and take-home monthly pay among respondents missing neither variable to determine the appropriate value for gross monthly pay. In particular, a random draw was selected from the observed distribution of relative taxes, where "relative tax" is defined as the proportion of a respondent's pay devoted to taxes. We then used the randomly drawn relative tax to determine an imputed gross monthly pay for four cases with missing data for C_MainCurJobMnthPay_i. As noted in Table IV.3, we applied hot-deck imputations to only four of the jobs variables: (1) C1_i, (2) C2_1_1d_i, (3) C8_1_i, and (4) C_TotCurMnthPay_i. For these variables, we used the level of education as a classing variable as well as additional classing and sorting variables specific to each variable, including a condition code variable for all but C_TotCurMnthPay_i.

Some of the variables in the above table had missing values that were not directly imputed. Rather, constituent variables not included in the table had missing values that were imputed and then combined to form the variables in the table. For example, we constructed C_TotCurWkHrs_i from the number of hours per week usually worked at the current main job plus the number of hours for each of the respondent's other jobs. In most cases, the respondent worked one job, so we set C_TotCurWkHrs_i equal to C8_1_i. However, if the respondent worked more than one job and the number of hours in secondary jobs was imputed, we constructed C_TotCurWkHrs_i from imputed variables.

Table IV.3. Current Jobs Imputations

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
C1_i	Count of current jobs	1 logical, 2 hot deck	3	445	0.67
C2_1_1d_i	Main current job SOC code to one digit	1 hot deck[a]	1	445	0.22
C8_1_i	Hours per week usually worked at current main job	19 hot deck,[b] 2 imputed by distributional assumptions	21	445	4.72
C_TotCurWkHrs_i	Total weekly hours at all current jobs	19 hot deck,[c] 5 constructed from imputed variables	24	445	5.39
C_TotCurHrMnth_i	Total hours per month at all current jobs	24 constructed from imputed variables	24	445	5.39
C_MainCurJobHrPay_i	Hourly pay at current main job	1 logical, 69 constructed from imputed variables	70	445	15.73
C_MainCurJobMnthPay_i	Monthly pay at current main job	12 logical, 4 imputed by distributional assumptions, 65 constructed from imputed variables	81	445	18.20
C_TotCurMnthPay_i	Total monthly salary all current jobs	15 logical, 65 hot deck, 7 constructed from imputed variables	87	445	19.55

Source: NBS–General Waves Round 5.

Note: The "number missing" and "number eligible" counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The "number missing" is a count of item nonrespondents, and the "number eligible" includes both item respondents and item nonrespondents. The "percentage imputed" is the "number missing" divided by the "number eligible", and is unweighted.

[a]Imputations for current job variables excluded two cases coded as "don't know" or "refused" in Item B24, which were imputed as currently not working in Item B24_i. Imputations for current job variables include another case coded as "don't know or "refused" in Item B24 that was imputed as currently working in item B24_i.

[b]Imputations for current job variables excluded two cases coded as "don't know" or "refused" in Item B24, which were imputed as currently not working in Item B24_i. Imputations for current job variables include another case coded as "don't know or "refused" in Item B24 that was imputed as currently working in Item B24_i.

[c]If C8_1_i was imputed by hot deck and the respondent had only one job, the flag indicated that C_TotCurWkHrs_i was imputed by hot deck, even though the variable was not processed in the hot-deck program.

4. Section I: Health Status Variables

Section I of the NBS—General Waves accounted for 57 health status variables in which imputations were applied. Tables IV.4 and IV.5 identify the 57 imputed variables and the methods of imputation used for each variable. The items cover a range of topics, from the respondent's general health to specific questions on instrumental activities of daily living (IADLs), activities of daily living (ADLs), and other health and coping indicators. A series of questions pertaining to the respondent's use of illicit drugs and alcohol is also included in Section I.

Table IV.4. Health Status Imputations, Questionnaire Variables

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
I1_i	Health during the past four weeks	12 hot deck	10	4,062	0.30
I9_i	Current health	25 hot deck	25	4,062	0.62
I17b_i	Blind or difficulty seeing, even with glasses	1 logical, 29 hot deck	30	4,062	0.73
I19_i	Uses special equipment because of difficulty seeing	24 logical, 6 hot deck	30	789	3.80
I21_i	Deaf or difficulty hearing	2 logical, 26 hot deck	28	4,062	0.69
I22_i	Able to hear normal conversation at all	22 logical, 17 hot deck	39	500	7.80
I23_i	Uses special equipment because of difficulty hearing	22 logical, 3 hot deck	25	500	5.00
I25_i	Difficulty having speech understood	3 logical, 30 hot deck	33	4,062	0.81
I26_i	Able to have speech understood at all	22 logical, 13 hot deck	35	1,185	2.95
I27_i	Uses special equipment because of difficulty speaking	22 logical, 5 hot deck	27	1,185	2.28
I29_i	Difficulty walking or climbing stairs without assistance	2 logical, 24 hot deck	26	4,062	0.64
I30_i	Able to walk without assistance at all	13 logical, 19 hot deck	32	2,155	1.48
I31_i	Uses special equipment because of difficulty walking	13 logical, 12 hot deck	25	2,155	1.16
I34_i	Able to climb stairs at all	13 logical, 20 hot deck	33	2,155	1.53
I35_i	Difficulty lifting and carrying 10 pounds	3 logical, 32 hot deck	35	4,062	0.86
I36_i	Able to lift or carry 10 pounds at all	16 logical, 49 hot deck	65	1,911	3.40
I37_i	Difficulty using hands or fingers	1 logical, 24 hot deck	25	4,062	0.61
I38_i	Able to use hands or fingers at all	17 logical, 19 hot deck	36	1,107	3.25
I39_i	Difficulty reaching over head	2 logical, 40 hot deck	42	4,062	1.03
I40_i	Able to reach over head at all	27 logical, 17 hot deck	44	1,165	3.78
I41_i	Difficulty standing	48 hot deck	48	4,062	1.18

TABLE IV.4 (continued)

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
I42_i	Able to stand at all	23 logical, 22 hot deck	45	2,476	1.82
I43_i	Difficulty stooping	1 logical, 44 hot deck	45	4,062	1.11
I44_i	Able to stoop at all	19 logical, 40 hot deck	59	2,398	2.46
I45_i	Difficulty getting around inside home	27 hot deck	27	4,062	0.66
I46_i	Needs help to get around inside home	23 logical, 14 hot deck	37	683	5.42
I47_i	Difficulty doing errands alone	7 logical, 38 hot deck	45	4,062	1.11
I48_i	Needs help to get around outside home	19 logical, 33 hot deck	52	2,312	2.25
I49_i	Difficulty getting into/out of bed	1 logical, 36 hot deck	37	4,062	0.91
I50_i	Needs help getting into/out of bed	25 logical, 18 hot deck	43	1,137	3.78
I51_i	Difficulty bathing or dressing	6 logical, 40 hot deck	46	4,062	1.13
I52_i	Needs help bathing or dressing	31 logical, 13 hot deck	44	1,121	3.93
I53_i	Difficulty shopping	15 logical, 41 hot deck	56	4,062	1.38
I54_i	Needs help shopping	27 logical, 20 hot deck	47	1,501	3.13
I55_i	Difficulty preparing own meals	6 logical, 33 hot deck	39	4,062	0.96
I56_i	Needs help to prepare meals	18 logical, 25 hot deck	43	1,594	2.70
I57_i	Difficulty eating	1 logical, 29 hot deck	30	4,062	0.73
I58_i	Needs help to eat	26 logical, 2 hot deck	28	562	4.98
I59_i	Trouble concentrating or remembering	51 hot deck	51	4,062	1.26
I60_i	Trouble coping with stress	76 hot deck	76	4,062	1.87
I61_i	Trouble getting along with people	66 hot deck	66	4,062	1.62
CageScore_Indicator_i	CAGE Alcohol Score	36 constructed from imputed variables	36	4,062	0.89
I72_i	Uses drugs in larger amounts than prescribed	48 hot deck	48	4,062	1.18

Source: NBS–General Waves Round 5.

Note: The “number missing” and “number eligible” counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The “number missing” is a count of item nonrespondents, and the “number eligible” includes both item respondents and item nonrespondents. The “percentage imputed” is the “number missing” divided by the “number eligible”, and is unweighted.

Table IV.5. Health Status Imputations, Constructed Variables

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
C_EquipFuncLim_I	Uses equipment/device for functional/sensory limitation	20 constructed from imputed variables	20	4,062	0.49
C_NumSenLim_i	Number of sensory limitations	55 constructed from imputed variables	55	4,062	1.35
C_NumSevSenLim_i	Number of severe sensory limitations	47 constructed from imputed variables	47	4,062	1.16
C_NumPhyLim_i	Number of physical functional limitations	102 constructed from imputed variables	102	4,062	2.51
C_NumSevPhyLim_i	Number of severe physical functional limitations	141 constructed from imputed variables	141	4,062	3.47
C_NumEmotLim_i	Number of emotional/social limitations	133 constructed from imputed variables	133	4,062	3.27
C_NumADLs_i	Number of impaired ADL	65 constructed from imputed variables	65	4,062	1.60
C_NumADLAssist_i	Number of ADL requiring assistance	55 constructed from imputed variables	55	4,062	1.35
C_NumIADLs_i	Number of IADL difficulties	73 constructed from imputed variables	73	4,062	1.80
C_NumIADLAssist_i	Number of IADL requiring assistance	71 constructed from imputed variables	71	4,062	1.75
C_PCS8TOT_i	Physical summary score	193 constructed from imputed variables	193	4,062	4.75
C_MCS8TOT_i	Mental summary score	193 constructed from imputed variables	193	4,062	4.75
C_DrugDep_i	Drug dependence	48 constructed from imputed variables	48	4,062	1.18

Source: NBS–General Waves Round 5.

Note: The “number missing” and “number eligible” counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The “number missing” is a count of item nonrespondents, and the “number eligible” includes both item respondents and item nonrespondents. The “percentage imputed” is the “number missing” divided by the “number eligible”, and is unweighted.

The following is an example of a logical assignment in Section I: If respondents did not answer whether they were blind or experienced difficulty seeing even when wearing glasses or contact lenses (Item I17b), but indicated that they required special devices to see because they had difficulty seeing (Item I19), then we logically assigned “yes” to Item I17b_i.

As in previous sections, “constructed from imputed variables” refers to the fact that we imputed the constituent variables of each constructed variable. The only classing variable common to all imputations was the code variable for the collapsed condition. We also used age and gender in most imputations. The other classing and sorting variables were specific to the variable being imputed.

5. Section K: Sources of Income Other Than Employment

The imputed variables in Section K are constructed variables that pertain to nonemployment-based income and include workers' compensation, private disability claims, unemployment, and other sources of regular income, as described in Table IV.6

Table IV.6. Imputations on Sources of Income Other Than Employment

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
C_AmtPrivDis_i	Amount received from private disability last month	107 logical, 17 imputed by descriptive statistics using specialized procedures	124	4,062	3.05
C_AmtWorkComp_i	Amount received from workers' compensation last month	50 logical, 2 imputed by descriptive statistics using specialized procedures	52	4,062	1.28
C_AmtVetBen_i	Amount received from veterans' benefits last month	43 logical, 16 imputed by descriptive statistics using specialized procedures	59	4,062	1.45
C_AmtPubAssis_i	Amount received from public assistance last month	58 logical, 14 imputed by descriptive statistics using specialized procedures	72	4,062	1.77
C_AmtUnemply_i	Amount received from unemployment benefits last month	43 logical, 2 imputed by descriptive statistics using specialized procedures	45	4,062	1.11
C_AmtPrivPen_i	Amount received from private pension last month	55 logical, 9 imputed by descriptive statistics using specialized procedures	64	4,062	1.57
C_AmtOthReg_i	Amount received from other regular sources last month	49 logical, 9 imputed by descriptive statistics using specialized procedures	58	4,062	1.43

Source: NBS–General Waves Round 5.

Note: The "number missing" and "number eligible" counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The "number missing" is a count of item nonrespondents, and the "number eligible" includes both item respondents and item nonrespondents. The "percentage imputed" is the "number missing" divided by the "number eligible", and is unweighted.

Items in Section K first asked respondents if they received money from a specific source and then asked for the specific amount received from that source. If a respondent could not provide a specific value, he or she answered a series of questions about whether the amount was above or below specific values. Respondents also had the option of providing a range of values, in which the options depended upon responses to a series of questions. After we classified the response according to a range of values provided by the respondent, we assigned the respondent the median of the specific values provided by others who gave responses within the same range. If a respondent could not say whether the actual value was above or below a specific threshold, we

first imputed the range (using random assignment), then assigned the median of the values provided by respondents who listed specific values within that range. If the respondent did not know if he or she received funds from a source, we used hot-deck imputation to determine whether such was the case and then proceeded as above.

The logical assignments in Section K derive from imputed values in the constituent questions. For example, Item K6 in the questionnaire asks whether the respondent received income from a variety of sources, and Item K7 asks the amount from each source for which a "yes" response was given. The first source listed (Item K6a) is private disability insurance. If the respondent was imputed not to have received private disability insurance (K6a_i), then the constructed variable C_AmtPrivDis_i (based on Item K7) was logically assigned "no." Otherwise, if any income was derived from private disability insurance but an imputation was required at some point in the sequence (either everything or just the individual's income was imputed), then the imputation flag indicated imputation by "special procedures."

For variables requiring hot-deck imputation, the classing variables were the same for all variables: an indicator of whether the respondent was a recipient of SSI, SSDI, or both; living situation; and education. Table IV.6 lists none of the variables requiring hot-deck imputation because they were just component variables for the delivered variables listed in the table.

6. Section L: Personal and Household Characteristics

We discussed race and ethnicity, derived from items L1 and L2 in the questionnaire, in Section 1 of this chapter. Other imputed variables that are personal and household characteristics also come from Section L. The questions from which the imputed variables were derived ask about education (L3_i), marital status (L8_i), cohabitation status (C_Cohab_i), number of children in household (C_NumChildHH_i), household size (C_Hhsize_i), and weight and height, which were used to derive body mass index (C_BMI_cat_i). Most of these variables were imputed early in imputation processing and were used in the imputation of variables imputed later in processing. Household income questions are also asked in Section L, which, in combination with C_Hhsize_i and C_NumChildHH_i, we use to derive the federal poverty level variable.

The imputation of poverty level required the imputation of annual income and household size. The annual income question was another case that required a specific value. If the respondent could not provide a specific value, he or she was asked if annual income fell within certain ranges. Some respondents provided a specific value, some provided a range of values, and some refused to provide any information. Although annual income was a key variable used in the imputation of poverty level, it was not included in Table IV.7 because it was not released in the final file. All missing values in C_FedPovertyLevel_cat1[36] were derived from the imputed annual incomes; hence, all missing values are "constructed from imputed variables." In Table IV.7, we identify the imputed variables in Section L.

[36] The name of this variable reflects the fact that the final variable was a categorical (as opposed to a continuous) measure of poverty level.

Logical assignments in Section L are based on related variables also in Section L. For example, a logical assignment for L11_i (living situation of beneficiary) would occur if the respondent did not answer Item L11 but indicated in Item L16 (number of adults in household) that only one adult lived in the household and indicated in Item L17 (number in household under 18 years old) the number of children living in the household. In this case, the value for L11_i would be logically assigned to 1 (lives alone) or 2 (lives with parent, spouse, or children), depending upon the response to Item L17.

The only classing variable common to all imputations for the variables listed in Table IV.7 was the collapsed condition code variable. Other classing and sorting variables were specific to the variable being imputed.

Table IV.7. Imputations of Personal and Household Characteristics

Variable Name	Description	Imputation Method	Number Missing	Number Eligible	Percentage Imputed
C_BMI_cat_i	Body mass index categories	1 logical, 190 hot deck	191	4,062	4.70
L3_i	Highest year/grade completed in school	99 hot deck	99	4,062	2.44
L8_i	Marital status	51 hot deck	51	4,062	1.26
L11_i	Living arrangements	4 logical, 51 hot deck	55	4,062	1.35
C_NumChildHH_i	Number of children living in household	1 logical, 28 hot deck, 16 constructed from imputed variables	45	4,062	1.10
C_HHsize_i	Household size	64 hot deck, 11 constructed from imputed variables	75	4,062	1.85
C_Cohab_i	Cohabitation status	2 logical, 49 hot deck	51	4,062	1.26
C_FedPovertyLevel_cat	2014 Federal poverty level	1,476 constructed from imputed variables	1,476	4,062	36.34

Source: NBS–General Waves Round 5.

Note: The “number missing” and “number eligible” counts exclude those who skipped out of the relevant question(s) based upon computer skip patterns. The “number missing” is a count of item nonrespondents, and the “number eligible” includes both item respondents and item nonrespondents. The “percentage imputed” is the “number missing” divided by the “number eligible”, and is unweighted.

V. ESTIMATING SAMPLING VARIANCE

The sampling variance of an estimate derived from survey data for a statistic (such as a total, a mean or proportion, or a regression coefficient) is a measure of the random variation among estimates of the same statistic computed over repeated implementation of the same sample design with the same sample size on the same population. The sampling variance is a function of the population characteristics, the form of the statistic, and the nature of the sampling design. The two general forms of statistics are linear combinations of the survey data (for example, a total) and nonlinear combinations. The latter include the ratio of two estimates (for example, a mean or proportion in which both the numerator and denominator are estimated) and more complex combinations, such as regression coefficients. For linear estimates with simple sample designs (such as a stratified or unstratified simple random sample) or complex designs (such as stratified multistage designs), explicit equations are available to compute the sampling variance. For the more common nonlinear estimates with simple or complex sample designs, explicit equations generally are not available, and various approximations or computational algorithms provide an essentially unbiased estimate of the sampling variance.

The NBS—General Waves sample design involves stratification and unequal probabilities of selection. Variance estimates calculated from NBS—General Waves data must incorporate the sample design features to obtain the correct estimate. Most procedures in standard statistical packages, such as SAS, STATA, and SPSS, are not appropriate for analyzing data from complex survey designs, such as the NBS—General Waves design. These procedures assume independent, identically distributed observations or simple random sampling with replacement. Although the simple random sample variance may approximate the true sampling variance for some surveys, it likely underestimates substantially the sampling variance with a design as complex as that used for the NBS—General Waves. Complex sample designs have led to the development of a variety of software options that require the user to identify essential design variables such as strata, clusters, and weights.[37]

The most appropriate sampling variance estimators for complex sample designs such as the NBS—General Waves are the procedures based on the Taylor series linearization of the nonlinear estimator that use explicit sampling variance equations and procedures based on forming pseudo-replications[38] of the sample. The Taylor series linearization procedure is based on a classic statistical method in which a nonlinear statistic may be approximated by a linear combination of the components within the statistic. The accuracy of the approximation depends upon the sample size and the complexity of the statistic. For most commonly used nonlinear statistics (such as ratios, means, proportions, and regression coefficients), the linearized form has

[37] A web site that reviews software for variance estimation from complex surveys, created with the encouragement of the Section on Survey Research Methods of the American Statistical Association, is available at http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. The site lists software packages available for personal computers and provides direct links to the home pages of the packages. The site also contains articles and links to articles that provide general information about variance estimation as well as links to articles that compare features of the software packages.

[38] Pseudo-replications of a specific survey sample, as opposed to true replications of the sampling design, involve the selection of several independent subsamples from the original sample data with the same sampling design. The subsamples may be random (as in a bootstrap) or restricted (as in balanced repeated replication).

been developed and has good statistical properties. Once a linearized form of an estimate is developed, the explicit equations for linear estimates may be used to estimate the sampling variance. The sampling variance may be estimated by using many features of the sampling design (for example, finite population corrections, stratification, multiple stages of selection, and unequal selection rates within strata). This is the basic variance estimation procedure used in all SUDAAN procedures as well as in the survey procedures in SAS, STATA, and other software packages that accommodate simple and complex sampling designs. To calculate the variance, sample design information (such as stratum, analysis weight, and so on) is needed for each sample unit.

Currently, several survey data analysis software packages use the Taylor series linearization procedure and explicit sampling variance equations. Therefore, we developed the variance estimation specifications needed for the Taylor series linearization (PseudoStrata and PseudoPSU). Appendix E provides example code for the procedure with SAS and the survey data analysis software SUDAAN.[39] Details about SAS syntax are available from the SAS Institute (2015). Details about SUDAAN syntax are available from RTI International (Research Triangle Institute 2014).

[39] The example code provided in Appendix E is for simple descriptive statistics using the procedures DESCRIPT in SUDAAN and SURVEYMEANS in SAS. Other procedures in SAS (SURVEYREG, SURVEYFREQ, and SURVEYLOGISTIC) and in SUDAAN (CROSSTAB, REGRESS, LOGISTIC, MULTILOG, LOGLINK, and SURVIVAL) are available for complex analyses. Given that SUDAAN was created specifically for survey data, the range of analyses that may be performed with these data in SUDAAN is much wider than that in SAS.

REFERENCES

Agresti, A. Categorical Data Analysis. New York: John Wiley and Sons, 1990.

Akaike, H. "A New Look at the Statistical Model Identification." *IEEE Transaction on Automatic Control*, AC-19, 1974, pp. 716-723.

Barrett, K., D. Wright, and G. Livermore. "The National Beneficiary Survey-General Waves: Round 5 Questionnaire." Washington, DC: Mathematica Policy Research, 2016.

Biggs, D., B. deVille, and E. Suen. "A Method of Choosing Multiway Partitions for Classification and Decision Trees." *Journal of Applied Statistics*, vol. 18, 1991, pp. 49-62.

Bush, C., R. Callahan, and J. Markesich. "The National Beneficiary Survey—General Waves: Round 5 Public-Use File Codebook." Washington, DC: Mathematica Policy Research, 2017.

Bush, C., R. Callahan, and J. Markesich. "The National Beneficiary Survey—General Waves: Round 5 Restricted-Use File Codebook." Washington, DC: Mathematica Policy Research, 2017.

Cox, D.R., and E.J. Snell. The Analysis of Binary Data, Second Edition. London: Chapman and Hall, 1989.

Folsom, R., F. Potter, and S. Williams. "Notes on a Composite Size Measure for Self to Weighting Samples in Multiple Domains." *Proceedings of the American Statistical Association, Section on Survey Research Methods*, 1987, pp. 792-796.

Grau, E. "The National Beneficiary Survey-General Waves: Round 5: Nonresponse Bias Analysis." Washington, DC: Mathematica Policy Research, 2017.

Hosmer, D.W., Jr., and S. Lemeshow. "Goodness-of-Fit Tests for the Multiple Logistic Regression Model. *Communications in Statistics, Theory and Methods*, vol. A9, no. 10, 1980, pp. 1043-1069.

Kass, G.V. "An Exploratory Technique for Investigating Large Quantities of Categorical Data." *Applied Statistics*, vol. 29, 1980, pp. 119-127.

Magidson, J. SPSS for Windows CHAID Release 6.0. Belmont, MA: Statistical Innovations, Inc., 1993.

NAACCR Expert Panel on Hispanic Identification. "Report of the NAACCR Expert Panel on Hispanic Identification 2003." Springfield, IL: North American Association of Central Cancer Registries, 2003.

Research Triangle Institute. *SUDAAN Language Manual, Release 9.0*. Research Triangle Park, NC: Research Triangle Institute, 2014.

SAS® Institute. SAS/STAT® 9.1 User's Guide. Cary, NC: SAS Institute, 2015.

Skidmore, S., D. Wright, Barrett, K and E. Grau. "National Beneficiary Survey—General Waves Round 5 (volume 2 of 3): Data Cleaning and Identification of Data Problems." Washington, DC: Mathematica Policy Research, 2017.

U.S. Department of Education. National Center for Education Statistics. "A Study of Imputation Algorithms." Working Paper No. 2001-17. Ming-xiu Hu and Sameena Salvucci. Washington, DC. 2001.

Wright, D., K. Barrett, S. Skidmore, E. Grau, Y. Zheng, K. Barrett, C. Bush and J. Markesich. "The National Beneficiary Survey-General Waves: Round 5 (Volume 3 of 3): User's Guide for Restricted and Public Use Data Files." Washington, DC: Mathematica Policy Research, 2017.

APPENDIX A

OTHER SPECIFY AND OPEN-ENDED ITEMS WITH ADDITIONAL CATEGORIES CREATED DURING CODING

This page has been left blank for double-sided copying.

Appendix A. "Other/Specify" and Open-Ended Items with Additional Categories Created During Coding

Question #	Question Text	Current Response Options	Additional Categories Created
B25	What are they (the other reasons you are not working that I didn't mention)?	a = A physical or mental condition prevents [you/him/her] from working b = [You/NAME] cannot find a job that [you are/(he/she) is] qualified for c = [You do/NAME does] not have reliable transportation to and from work d = [You are/NAME is] caring for someone else. f = [You/NAME] cannot find a job [you want/(he/she) wants] g = [You are/NAME is] waiting to finish school or a training program. h = Workplaces are not accessible to people with [your/NAME's] disability. i = [You do/NAME does] not want to lose benefits such as disability, worker's compensation, or Medicaid j = [Your/NAME's] previous attempts to work have been discouraging l = Others do not think [you/NAME] can work m=Employers will not give [you/NAME] a chance to show that [you/he/she] can work. n = [You/NAME] does not have the special equipment or medical devices that [you/he/she] would need in order to work. o = [You/NAME] cannot get the personal assistance [you need/he needs/she needs] in order to get ready for work each day	p=Cannot find a job/job market is bad q=Lack skills
B29_6	What benefits [were/was] [you/NAME] most worried about losing?	1= Private disability insurance 2= Workers' compensation 3= Veterans' benefits 4= Medicare 5= Medicaid 6= SSA disability benefits 7= Public assistance or welfare 8= Food stamps 9= Personal assistance services (pas) 10= Unemployment benefits 11= Other state disability benefits 12= Other government programs 13= Other	14= Health insurance unspecified

Question #	Question Text	Current Response Options	Additional Categories Created
B29_10	What benefits [were/was] [you/NAME] most worried about losing?	01= Private Disability Insurance 02= Workers' compensation 03= Veterans' benefits 04= Medicare 05= Medicaid 06= SSA Disability Benefits 07= Public Assistance or Welfare 08= Food Stamps 09= Personal Assistance Services (PAS) 10= Unemployment Benefits 11= Other State Disability Benefits 12= Other government programs 13= Other	14= Health insurance unspecified
B29_11b	What benefits [were/was] [you/NAME] most worried about losing?	01= Private Disability Insurance 02= Workers' compensation 03= Veterans' benefits 04= Medicare 05= Medicaid 06= SSA Disability Benefits 07= Public Assistance or Welfare 08= Food Stamps 09= Personal Assistance Services (PAS) 10= Unemployment Benefits 11= Other State Disability Benefits 12= Other government programs 13= Other	14= Health insurance unspecified
C35	Are there any changes in [your/NAME's] [main/current] job or workplace related to [your/his/her] mental or physical condition that [you need/he/she needs], but that have not been made? (IF YES) What are those changes?	<OPEN>	a= Need special equipment or assistive b= Need changes in [your/NAME's] work schedule c= Need changes to the tasks [you were/NAME was] assigned or how they are performed d= Need changes to the physical work environment e= Need co-workers or others to assist [you/NAME]? f=Need other changes

Question #	Question Text	Current Response Options	Additional Categories Created
C39b	[Do you/Does NAME] work fewer hours or earn less money than [you/he/she] could because [you/he/she]:	a = [Are/Is] taking care of children or others? b = [Are/Is] enrolled in school or a training program? c = Want[s] to keep Medicare or Medicaid coverage? d = Want[s] to keep cash benefits [you/he/she] need such as disability or workers' compensation? e = Just [do/does] not want to work more? f = Are there any reasons I didn't mention why [you are/NAME is] working or earning less than [you/he/she] could?	g=[Are/is] in poor health or [have/has] health concerns?
C39_2	What benefits have been reduced or ended as a result of [your/NAME's] (main/current) job?	01 = Private Disability Insurance 02 = Workers' compensation 03 = Veterans' benefits 04 = Medicare 05 = Medicaid 06 = SSA Disability Benefits 07 = Public Assistance or Welfare 08 = Food Stamps 09 = Personal Assistance Services (PAS) 10 = Unemployment Benefits 11 = Other State Disability Benefits 12 = Other government programs 13 = Other	14= Health insurance unspecified

Question #	Question Text	Current Response Options	Additional Categories Created
D23	Why did [you/NAME] stop working at this job?	LAYOFF, FIRED, RETIRED 1=LAYOFF, PLANT CLOSED 2=FIRED 3=RETIRED/OLD AGE 4=JOB WAS TEMPORARY AND ENDED PROBLEMS WITH JOB 5=DID NOT LIKE SUPERVISOR OR CO-WORKERS 6=DID NOT LIKE JOB DUTIES 7=DID NOT LIKE JOB EARNINGS 8=DID NOT LIKE BENEFITS 9=DID NOT LIKE OPPORTUNITIES FOR ADVANCEMENT 10=DID NOT LIKE LOCATION 11=DID NOT GET ACCOMMODATIONS THAT WERE NEEDED OTHER PROBLEMS 12=TRANSPORTATION PROBLEMS 13=DECIDED TO GO TO SCHOOL 14=CHILD CARE RESPONSIBILITIES (PREGNANT) 15=OTHER FAMILY OR PERSONAL REASONS DISABILITY 16=DISABILITY GOT WORSE 17=BECAME DISABLED 18=OTHER (SPECIFY: <OPEN>)	19= Moved to another area 20= Found another job 21= Loss or potential loss of government benefits 22= Work schedule
D25	Did you work fewer hours or earn less money than you could have because [you/he/she] you…	a= [Were/Was] taking care of somebody else? b= [Were/Was] enrolled in school or a training program? c= Wanted to keep Medicare or Medicaid coverage d= Wanted to keep cash benefits such as disability or workers compensation? e= Just didn't want to work more? f= Are there any reasons I didn't mention why [you/NAME] might have chosen to work or earn less than [you/he/she] could have during 2004? (SPECIFY: <OPEN>)	g=Had medical problems/complications

Question #	Question Text	Current Response Options	Additional Categories Created
D26	In 2014, do you think [you/NAME] could have worked or earned more if [you/he/she] had:	a=Help caring for [your/his/her] children or others in the household? b=Help with [your/his/her] own personal care such as bathing, dressing, preparing meals, and doing housework? c=Reliable transportation to and from work? d=Better job skills? e=A job with a flexible work schedule? f=Help with finding and getting a better job? g=Any special equipment or medical devices? (SPECIFY: <OPEN>) h=Is there anything else that I didn't mention that would have helped [you/NAME] to work or earn more during 2004? (SPECIFY: <OPEN>)	i=Better health/treatment j=More supportive/helpful employer and/or coworker
G7	Thinking about [PROVIDER FROM G2], was this place:	01=A state agency 02=A private business 03=Some other type of place? (SPECIFY: <OPEN>)	04=School
G18	Thinking about [NEW PROVIDER FROM G16], was this place:	01=A clinic, 02=A hospital, 03=A doctor's office, or 04=Some other type of place? (SPECIFY: <OPEN>)	05=A school 06=A nursing home/group home 07=A government agency 08=In home care 09=A medical equipment store 10=A rehabilitation/counseling center 11=Physical therapy center
G22	Thinking about [NEW PROVIDER FROM G20], was this place:	01=A mental health agency, 02=A clinic, 03=A hospital, 04=A doctor's office, or 05=Some other type of place? (SPECIFY: <OPEN>)	06=Residential treatment program/facility 07=Rehab center/counseling center/day program 08=Church or religious institution

Question #	Question Text	Current Response Options	Additional Categories Created
G36	In 2014, please tell me if [you/NAME] received any of the following services from [PROVIDER FROM G30_1 DE-DUPLICATED LIST IF USED IN 2004]. Did [you/he/she] receive:	a=Physical therapy? b=Occupational therapy? c=Speech therapy? e=Special equipment or devices? f=Personal counseling or therapy? g=Group therapy? d= Medical services? h=A work or job assessment? i=Help to find a job? j=Training to learn a new job or skill? k=Advice about modifying [your/his/her] job or work place? l=On-the-job training, job coaching, or support services? m=Anything else that I didn't mention? (SPECIFY: <OPEN>)	n=Scholarships/grants/loans 0=Prescription services/medication
G61	Why [were you/was NAME] unable to get these services?	<OPEN>	01= Not eligible/request refused 02= Lack information on how to get services/didn't know about services 03= Could not afford/insurance would not cover 04= Did not try to get services 05= Too difficult/too confusing to get services 06=Problems with the service or agency 07=Other
K14	What other assistance did [you/NAME] receive last month?	<OPEN>	01=Housing Assistance 02=Energy Assistance 03=Food assistance 04=Other
L12	The next question is about the place where you live. Was this place a…	01=Single family home? 02=Mobile home? 03=Regular apartment? 04=Supervised apartment? 05=Group home? 06=Halfway house? 07=Personal care or board and care home? 08=Assisted living facility? 09=Nursing or convalescent home? 10=Center for independent living? 11=Some other type of supervised group residence or facility? 12=Something else?	13=Homeless

APPENDIX B

SOC MAJOR AND MINOR OCCUPATION CLASSIFICATIONS

This page has been left blank for double-sided copying.

Appendix B. SOC Major and Minor Occupation Classifications

Code	Occupation
	Management
111	Top Executives
112	Advertising, Marketing, PR, Sales
113	Operations Specialist Managers
119	Other Management Occupations
	Business /Financial Operations
131	Business Operations Specialist
132	Financial Specialist
	Computer and Mathematical Science
151	Computer Specialist
152	Mathematical Science Occupations
	Architecture and Engineering
171	Architects, Surveyors and Cartographers
172	Engineers
173	Drafters, Engineering and Mapping Technicians
	Life, Physical and Social Science
191	Life Scientists
192	Physical Scientists
193	Social Scientists and Related Workers
194	Life, Physical and Social Science Technicians
	Community and Social Services
211	Counselors, Social Workers and Other Community and Social Service Specialists
212	Religious Workers
	Legal
231	Lawyers, Judges and Related Workers
232	Legal Support Workers
	Education, Training and Library
251	Postsecondary Teachers
252	Primary, Secondary and Special Education School Teachers
253	Other Teachers and Instructors
254	Librarians, Curators and Archivists
259	Other Education, Training and Library Occupations

Code	Occupation
	Arts, Design, Entertainment, Sports and Media
271	Art and Design Workers
272	Entertainers and Performers, Sports and Related Workers
273	Media and Communication Workers
274	Media and Communication Equipment Workers
	Healthcare Practitioner and Technical Occupations
291	Health Diagnosing and Treating Practitioners
292	Health Technologists and Technicians
299	Other Healthcare Practitioner and Technical Occupations
	Healthcare Support
311	Nursing, Psychiatric and Home Health Aides
312	Occupational and Physical Therapist Assistants and Aides
319	Other Healthcare Support Occupations
	Protective Service
331	Supervisors, Protective Service Workers
332	Firefighting and Prevention Workers
333	Law Enforcement Workers
339	Other Protective Service Workers
	Food Preparation and Serving Related
351	Supervisors, Food Preparation and Food Serving Workers
352	Cooks and Food Preparation Workers
353	Food and Beverage Serving Workers
359	Other Food Preparation and Serving Related Workers
	Building and Grounds Cleaning and Maintenance
371	Supervisors, Building and Grounds Cleaning and Maintenance Workers
372	Building Cleaning and Pest Control Workers
373	Grounds Maintenance Workers
	Personal Care and Service Occupations
391	Supervisors, Personal Care and Service Workers
392	Animal Care and Service Workers
393	Entertainment Attendants and Related Workers
394	Funeral Service Workers
395	Personal Appearance Workers
396	Baggage Porters, Bellhops, and Concierges
397	Tour and Travel Guides
399	Other Personal Care and Service Workers

Code	Occupation
	Sales and Related Occupations
411	Supervisors, Sales Workers
412	Retail Sales Workers
413	Sales Representative, Services
414	Sales Representative, Wholesale and Manufacturing
419	Other Sales and Related Workers
	Office and Administrative Support
431	Supervisors, Office and Administrative Support Workers
432	Communications Equipment Operators
433	Financial Clerks
434	Information and Record Clerks
435	Material Recording, Scheduling Dispatching, and Distribution Workers
436	Secretaries and Administrative Assistants
439	Other Office and Administrative Support Workers
	Farming, Fishing and Forestry Workers
451	Supervisors, Farming, Fishing and Forestry Workers
452	Agricultural Workers
453	Fishing and Hunting Workers
454	Forest, Conservation and Logging Workers
	Construction and Extraction Occupations
471	Supervisors, Construction and Extraction Workers
472	Construction Trade Workers
473	Helpers, Construction Trades
474	Other Construction and Related Workers
475	Extraction Workers
	Installation, Maintenance and Repair Occupations
491	Supervisors, Installation, Maintenance and Repair Workers
492	Electrical and Electronic Equipment Mechanics, Installers and Repairers
493	Vehicle and Mobile Equipment Mechanics, Installers and Repairers
494	Other Installation, Maintenance and Repair Occupations
	Production Occupations
511	Supervisors, Production Workers
512	Assemblers and Fabricators
513	Food Processing Workers
514	Metal Workers and Plastic Workers
515	Printing Workers
516	Textile, Apparel, and Furnishing Workers

Code	Occupation
517	Woodworkers
518	Plant and System Operators
519	Other Production Occupations
	Transportation and Material Moving Occupations
531	Supervisors, Transportation and Material Moving Workers
532	Air Transportation Workers
533	Motor Vehicle Operators
534	Rail Transportation Workers
535	Water Transportation Workers
536	Other Transportation Workers
537	Material Moving Workers
	Military Specific Occupations
551	Military Officer and Tactical Operations Leaders/Managers
552	First-Line Enlisted Military Supervisors/Managers
553	Military Enlisted Tactical Operations and Air/Weapons Specialists and Crew Members

APPENDIX C

NAICS INDUSTRY CODES

This page has been left blank for double-sided copying.

Appendix C. NAICS Industry Codes

Code	Description
11	**Agriculture, Forestry Fishing and Hunting**
111	Crop Production
112	Animal Production and Aquaculture
113	Forestry and Logging
114	Fishing, Hunting and Trapping
115	Support Activities for Agriculture and Forestry
21	**Mining, Quarrying, and Oil and Gas Extraction**
211	Oil and Gas Extraction
212	Mining (except Oil and Gas)
213	Support Activities for Mining
22	**Utilities**
221	Utilities
23	**Construction**
236	Construction of Buildings
237	Heavy and Civil Engineering Construction
238	Specialty Trade Contractors
31-33	**Manufacturing**
311	Food Manufacturing
312	Beverage and Tobacco Product Manufacturing
313	Textile Mills
314	Textile Product Mills
315	Apparel Manufacturing
316	Leather and Allied Product Manufacturing
321	Wood Product Manufacturing
322	Paper Manufacturing
323	Printing and Related Support Activities
324	Petroleum and Coal Products Manufacturing
325	Chemical Manufacturing
326	Plastics and Rubber Products Manufacturing
327	Nonmetallic Mineral Product Manufacturing
331	Primary Metal Manufacturing
332	Fabricated Metal Products Manufacturing
333	Machinery Manufacturing
334	Computer and Electronic Product Manufacturing
335	Electrical Equipment, Appliance and Component Manufacturing

Code	Description
336	Transportation Equipment Manufacturing
337	Furniture and Related Product Manufacturing
339	Miscellaneous Manufacturing
42	**Wholesale Trade**
423	Merchant Wholesalers, Durable Goods
424	Merchant Wholesalers, Nondurable Goods
425	Wholesale Electronic Markets and Agents and Brokers
44-45	**Retail Trade**
441	Motor Vehicle and Parts Dealers
442	Furniture and Home Furnishings Stores
443	Electronics and Appliance Stores
444	Building Material and Garden Equipment and Supplies Dealers
445	Food and Beverage Stores
446	Health and Personal Care Stores
447	Gasoline Stations
448	Clothing and Clothing Accessories Stores
451	Sporting Goods, Hobby, Musical Instrument, and Book Stores
452	General Merchandise Stores
453	Miscellaneous Store Retailers
454	Nonstore Retailers
48-49	**Transportation and Warehousing**
481	Air Transportation
482	Rail Transportation
483	Water Transportation
484	Truck Transportation
485	Transit and Ground Passenger Transportation
486	Pipeline Transportation
487	Scenic and Sightseeing Transportation
488	Support Activities for Transportation
491	Postal Service
492	Couriers and Messengers
493	Warehousing and Storage
51	**Information**
511	Publishing Industries (except Internet)
512	Motion Picture and Sound Recording Industries
515	Broadcasting (except Internet)

Code	Description
517	Telecommunications
518	Data Processing, Hosting, and Related Services
519	Other Information Services
52	**Finance and Insurance**
521	Monetary Authorities – Central Bank
522	Credit Intermediation and Related Activities
523	Securities, Commodity Contracts, and Other Financial Investments and Related Activities
524	Insurance Carriers and Related Activities
525	Funds, Trusts, and Other Financial Vehicles
53	**Real Estate and Rental and Leasing**
531	Real Estate
532	Rental and Leasing Services
533	Lessors of Nonfinancial Intangible Assets (except Copyrighted Works)
54	**Professional, Scientific, and Technical Services**
541	Professional, Scientific, and Technical Services
55	**Management of Companies and Enterprises**
551	Management of Companies and Enterprises
56	**Administrative and Supportive Waste Management and Remediation Services**
561	Administrative and Support Services
562	Waste Management and Remediation Services
61	**Educational Services**
611	Educational Services
62	**Health Care and Social Assistance**
621	Ambulatory Health Care Services
622	Hospitals
623	Nursing and Residential Care Facilities
624	Social Assistance
71	**Arts, Entertainment, and Recreation**
711	Performing Arts, Spectator Sports, and Related Industries
712	Museums, Historical Sites, and Similar Institutions
713	Amusement, Gambling, and Recreation Industries
72	**Accommodation and Food Services**
721	Accommodation
722	Food Services and Drinking Places

Code	Description
81	**Other Services (except Public Administration)**
811	Repair and Maintenance
812	Personal and Laundry Services
813	Religious, Grantmaking, Civic, Professional, and Similar Organizations
814	Private Households
92	**Public Administration**
921	Executive, Legislative, and Other General Government Support
922	Justice, Public Order, and Safety Activities
923	Administration of Human Resource Programs
924	Administration of Environmental Quality Programs
925	Administration of Housing Programs, Urban Planning, and Community Development
926	Administration of Economic Programs
927	Space Research and Technology
928	National Security and International Affairs

APPENDIX D

PARAMETER ESTIMATES AND STANDARD ERRORS FOR NONRESPONSE MODELS

This page has been left blank for double-sided copying.

Table D.1. Variables in the Location Logistic Propensity Model Representative Beneficiary Sample

Main Effects	Parameter Estimate[a]	Standard Error
Variables in the Beneficiary Location Model		
Count of addresses on file (MOVE)		
Only one address on file	1.221**	0.279
Two addresses on file	0.736**	0.183
Three addresses on file	0.431**	0.148
Four addresses on file	0.179	0.147
Five or more addresses on file, or no information	Ref. cell	
Count of phone numbers on file (PHONE)		
One to three phone numbers on file	-0.458†	0.315
Four or more phone numbers on file, or no information	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.796**	0.117
Age in range 30 to 39 years	-0.447**	0.116
Age in range 40 to 49 years	-0.251*	0.118
Age in range 50 to 64 years	Ref. cell	
Beneficiary's gender (GENDER)		
Male	-0.237*	0.111
Female	Ref. cell	
Indicator whether beneficiary and applicant for benefits are in same zip code (PDZIPSAME)		
Applicant and beneficiary live in different zip code	-0.019†	0.195
Applicant and beneficiary live in same zip code, or no information	Ref. cell	
Non-specialized economy county (CNTYNONSP)		
County's economy not dependent on farming, mining, manufacturing, government, or services	0.240	0.125
County that doesn't have this attribute	Ref. cell	
Two-Factor Interactions[b]		
PDZIPSAME*PHONE		
Indicator whether beneficiary and applicant for benefits are in same zip code, or no information * One to three phone numbers on file	0.755	0.335

[a] It is standard statistical practice to include main effects in models when they are a component of a significant interaction effect. Parameter estimates with a cross (†) represent such main effects that were included in the model for this reason. One star (*) and two stars (**) represent significance at the 5% and 1% levels respectively.

[b] All combinations for the listed interactions that are not shown are part of the reference cells.

Table D.2. Variables in the Cooperation Logistic Propensity Model Representative Beneficiary Sample

Main Effects	Parameter Estimate[a]	Standard Error
Variables in the Beneficiary Cooperation Model		
Beneficiary's age category (AGECAT)		
Age in range 30 to 39 years	-0.168*	0.081
Age in range 40 to 49 years	-0.170*	0.080
Age in range 18 to 29 years, or 50 to 64 years	Ref. cell	
Race of the beneficiary (RACE)		
White	-0.110	0.150
Not White or Unknown	Ref. cell	
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in nonmetropolitan area not adjacent to metropolitan area	0.569	0.293
Beneficiary resides in nonmetropolitan area adjacent to medium or small metropolitan area	0.419*	0.192
Beneficiary resides in nonmetropolitan area adjacent to large metropolitan area	0.603*	0.262
Beneficiary resides in metropolitan statistical area (MSA) of less than 250,000	0.219	0.161
Beneficiary resides in metropolitan statistical area (MSA) of 250,000-999,999	0.206†	0.167
Beneficiary resides in metropolitan statistical area (MSA) of 1 million or more	Ref. cell	
Beneficiary's gender (GENDER)		
Male	0.149	0.087
Female	Ref. cell	
Identity of payee relative to beneficiary (REPREPAYEE)		
Beneficiary received payments himself/herself	-0.842**	0.314
Beneficiary did not receive payments himself/herself, or unknown	Ref. cell	
Indicator whether beneficiary and applicant for benefits are in same zip code (PDZIPSAME)		
Applicant and beneficiary live in same zip code	-0.676†	0.419
Applicant and beneficiary live in different zip code	0.332†	0.207
No information	Ref. cell	
Count of phone numbers on file (PHONE)		
One phone number on file	0.307†	0.312
Two to six phone numbers on file	0.184†	0.188
More than six phone numbers on file, or unknown	Ref. cell	
Beneficiary's disability (DIG)		
Beneficiary has a cognitive disability	0.454	0.264
Beneficiary has a mental illness	0.597*	0.256
Beneficiary has a physical disability other than deafness	0.686**	0.255
Beneficiary is deaf, or information is unknown	Ref. cell	
Government-dependent economy county (CNTYGOV)		
County with a government-dependent economy	-0.350*	0.158
County that doesn't have this attribute	Ref. cell	

Main Effects	Parameter Estimate[a]	Standard Error
Service-dependent economy county (CNTYSVC)		
County with low levels of education	0.781†	0.305
County that doesn't have this attribute	Ref. cell	
County with poor quality/crowded housing (CNTYHSTRESS)		
County with poor quality/crowded housing	0.313†	0.251
County that doesn't have this attribute	Ref. cell	
County with high levels of persistent poverty (CNTYPERSPOV)		
County with high levels of persistent poverty	0.074†	0.263
County that doesn't have this attribute	Ref. cell	
County with low levels of education (CNTYLOWEDUC)		
County with low levels of education	0.370*	0.153
County that doesn't have this attribute	Ref. cell	
Two-Factor Interactions[b]		
CNTYHSTRESS*PDZIPSAME		
County with poor quality/crowded housing*Applicant & beneficiary live in same zip code	0.687***	0.187
Beneficiary missing one or both of these two attributes	Ref. cell	
CNTYHSTRESS*PHONE		
County with poor quality/crowded housing*One phone number on file	-0.401	0.365
County with poor quality/crowded housing*Two to six phone numbers on file	0.304	0.220
Beneficiary missing one or more of these attributes	Ref. cell	
CNTYHSTRESS*METRO		
County with poor quality/crowded housing*Metropolitan areas 250,000-999,999	-0.322	0.213
Beneficiary missing one or both of these two attributes	Ref. cell	
CNTYSVC*PHONE		
Service-dependent economy county*One phone number on file	0.018	0.364
Service-dependent economy county*Two to six phone numbers on file	0.456*	0.218
Beneficiary missing one or more of these attributes	Ref. cell	
CNTYPERSPOV*PDZIPSAME		
Persistent-poverty county*Applicant & beneficiary live in same zip code	-1.338**	0.435
Beneficiary missing one or both of these two attributes	Ref. cell	
CNTYSVC*PDZIPSAME		
Service-dependent economy county* Applicant & beneficiary live in same zip code	0.761***	0.268
Beneficiary missing one or both of these two attributes	Ref. cell	

[a] It is standard statistical practice to include main effects in models when they are a component of a significant interaction effect. Parameter estimates with a cross (†) represent such main effects that were included in the model for this reason.. One star (*) and two stars (**) represent significance at the 5% and 1% levels respectively.

[b]All combinations for the listed interactions that are not shown are part of the reference cells

This page has been left blank for double-sided copying.

APPENDIX E

SUDAAN PARAMETERS FOR NATIONAL ESTIMATES FROM THE NBS-GENERAL WAVES ROUND 5 SAMPLE

This page has been left blank for double-sided copying.

```
PROC DESCRIPT data="SASdatasetname" filetype=sas design=wr;
nest        A_STRATA A_PSU / missunit;
weight      "weight variable" ;
subpopn     "response variable" = "complete";
var         "analysis variables" ;
print nsum wsum mean semean deffmean / style=nchs
wsumfmt=f10.0 meanfmt=f8.4 semeanfmt=f8.4 deffmeanfmt=f8.4;
title "TTW National Estimates";
```

WEIGHT VARIABLES USED FOR CROSS-SECTIONAL ESTIMATES

Wtr5_ben

NEST VARIABLES USED FOR CROSS-SECTIONAL ESTIMATES

A_STRATA

a. A_STRATA = 1000 for non-certainty PSUs
b. A_STRATA = 2000 for Los Angeles County certainty PSU
c. A_STRATA = 3000 for Cook County certainty PSU

A_PSU

A_PSU=FIPSCODE-derived identifier for PSU or, in Los Angeles or Cook county, SSU

NOTES

1. Before each SUDAAN procedure, sort by A_STRATA and A_PSU

2. Use SUDAAN's SUBPOPN statement to define population for which estimates are wanted.

   For example, for estimates of SSI participant population, use SUBPOPN to define `SSI participants.**DOC**

This page has been left blank for double-sided copying.

This page has been left blank for double-sided copying.

www.mathematica-mpr.com

Improving public well-being by conducting high quality, objective research and data collection

PRINCETON, NJ ■ ANN ARBOR, MI ■ CAMBRIDGE, MA ■ CHICAGO, IL ■ OAKLAND, CA ■ SEATTLE, WA ■ TUCSON, AZ ■ WASHINGTON, DC ■ WOODLAWN, MD

MATHEMATICA
Policy Research