Mathematica
Progress Together

Final Report

National Beneficiary Survey-General Waves Round 6 (Volume 1 of 3): Editing, Coding, Imputation, and Weighting Procedures

November 14, 2019

Eric Grau, Yuhong Zheng, Susanna Vogel, Bevin Mory, Kim McDonald, Ryan Callahan, Hanzhi Zhou, Jason Markesich

Submitted to:

Social Security Administration
Office of Retirement and Disability Policy
ITC Building
500 E Street, SW, 9th Floor
Washington, DC 20254
Project Officer: Mark Trapani
Contract Number: 0600-12-60094

Submitted by:

Mathematica
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: Jason Markesich
Reference Number: 40160.224

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

CONTENTS

TABLES

ACRONYMS

AIC	Akaike's Information Criterion
AHRF	Area Health Resource File
CAPI	Computer-assisted personal interviewing
CATI	Computer-assisted telephone interviewing
CHAID	Chi-Squared Automatic Interaction Detector
DCF	Disability Control File
FRA	Full Retirement Age
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 publicly available reports are available from SSA on their website (https://www.ssa.gov/disabilityresearch/nbs_round_6.html):

- **User's Guide for Restricted- and Public-Use Data Files** (Callahan et al. 2019). 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 data collection for the National Beneficiary Survey (NBS)–General Waves. 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. 2019). 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** (Callahan et al. 2019). This document contains all items on Round 6 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). This report summarizes the editing, coding, imputation, and weighting procedures as well as the development of standard errors for Round 6 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 both the representative beneficiary and successful worker samples; 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** (McDonald et al. 2019). This report describes the data processing procedures performed for Round 6 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. 2019). This report discusses whether the nonresponse adjustments applied to the sampling weights of Round 6 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** (McDonald et al. 2019). 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 restricted-access file. The codebook also includes frequency distributions and means as appropriate.

I. INTRODUCTION

Sponsored by the Social Security Administration's (SSA's) Office of Retirement and Disability Policy, the National Beneficiary Survey (NBS)-General Waves collects data on the employment-related activities of working-age beneficiaries of Social Security Disability Insurance (SSDI) and Supplemental Security Income (SSI). In 2017, Mathematica Policy Research conducted the sixth round of data collection since the NBS began in 2004. We will implement a seventh round in 2019. The first four rounds of the survey—in 2004, 2005, 2006, and 2010—helped glean information about beneficiary impairments; health; living arrangements; family structure; occupation before disability; and use of non-SSA programs (for example, the Supplemental Nutrition Assistance Program, or SNAP). Rounds 1 to 4 also evaluated the Ticket to Work and Self-Sufficiency (TTW) program. In Rounds 5 to 7, we seek to uncover important information about the factors that promote beneficiaries' self-sufficiency and, conversely, the factors that impede beneficiaries' efforts to maintain employment.[1]

For Round 6 of the NBS, we met the goals of the study through two samples: (1) a sample of all beneficiaries (the representative beneficiary sample, or RBS), and (2) a sample of a subset of beneficiaries who maintained a minimum level of earnings for a sustained period (a successful worker sample, or SWS). The survey was administered to both of these cross-sectional samples simultaneously; a subset of SWS cases will be followed longitudinally in Round 7. Mathematica collected data by using computer-assisted telephone interviewing (CATI). We deployed in-person field locators to follow-up with most CATI non-respondents,[2] then conducted the interviews via CATI with a cell phone provided by the field locator. Computer-assisted personal interviewing (CAPI) was conducted only with sample members who preferred or needed an in-person interview to accommodate their disabilities. Both CATI and CAPI modes were fully integrated to simplify reporting and data processing.

In this report, we document the editing, coding, weighting, and imputation procedures, as well as the development of standard errors, for Round 6 of the NBS–General Waves. In Chapter II, we provide an overview of the variable naming, editing and coding, and construction conventions that were used in the data files and accompanying codebooks. In Chapter III, we discuss how we calculated the final analysis weights for the RBS and SWS, and the composite weights that combined weights from the two samples. In particular, we discuss how we calculated the initial sampling weights, adjusted them to account for nonresponse, post-stratified them to frame totals, and trimmed outlier weights when necessary. 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

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

[2] We did not employ field follow-up for a portion of the SWS. This portion, referred to as the "unclustered" sample, is described later in this chapter.

Chapter III. Finally, in Appendix E, we present SUDAAN and SAS parameters that could be used to generate national estimates from the Round 6 sample.[3]

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?
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?

SSA combines data from the NBS 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 estimates from the survey for other policymaking and program planning efforts.

We addressed the core research questions in Rounds 1 through 4 through two surveys, one of all beneficiaries (the RBS) and one of successful workers in the TTW program (the Ticket Participant Sample, or TPS). The NBS–General Waves (Rounds 5 through 7) no longer focuses on TTW. The survey design for Rounds 5 through 7 initially called for three national cross-sectional surveys of SSI and SSDI beneficiaries (the RBS)—one each in 2014, 2016, and 2018. It also called for cross-sectional surveys, in the same years, of beneficiaries whose benefits were suspended or terminated due to work (with a subset followed longitudinally across rounds). However, due to difficulties in identifying beneficiaries experiencing benefit suspense in SSA's administrative data, we subsequently revised the design to focus on beneficiaries with successful work attempts (the SWS). We delayed the start of NBS–General Waves by one year (from 2014, 2016, and 2018, to 2015, 2017, and 2019) to allow time to redesign the successful worker portion of the survey and sample, and we ultimately opted not to administer the SWS in Round 5. In Round 6, we conducted the second cross-sectional survey for the RBS in the NBS–General Waves,[4] using the same primary sampling units (PSUs) that were selected in Round 5, simultaneously conducting the first cross-sectional survey for the SWS. Some of the sampled

[3] SUDAAN and SAS are statistical packages that are used to analyze data. SAS is a general purpose package that includes procedures for survey data; SUDAAN was developed specifically for survey data. Details about SUDAAN are available in the SUDAAN user's manual (RTI, 2014)

[4] Although this is the second RBS in the NBS–General Waves, it is the sixth RBS over the entire history of the NBS.

SWS members will be followed in a longitudinal sample in Round 7.[5] A summary of the samples that were processed in Rounds 1 through 6, and will be processed in Round 7, is given in Table I.1.

Table I.1. Summary of samples processed in Rounds 1 through 7

Round	Year	Study	RBS	TPS	SWS
1	2004	NBS-TTW	√	√	
2	2005	NBS-TTW	√	√	
3	2006	NBS-TTW	√	√	
4	2010	NBS-TTW	√	√	
5	2015	NBS-General Waves	√		
6	2017	NBS-General Waves	√		√
7	2019	NBS-General Waves	√		√

B. NBS–General Waves sample design overview

For all survey rounds, the NBS has used a multistage sampling design. In Round 6, we used such a design for both the RBS and SWS, with an independently drawn supplemental single-stage sample for some subset populations.[6] We drew the SWS and RBS independently, from separate frames, although the SWS frame was a subset of the RBS frame.[7] This means that some sample members could have been selected for both the RBS and the SWS—which occurred for 91 individuals (of which, 38 responded). Because most analyses do not require combining the samples, we did not adjust the RBS and SWS weights for these duplicates. However, in the event that an analysis would require combining the samples, we also created composite weights that accounted for duplicates (individuals who were selected for both samples). These composite weights also accounted for those in the RBS that were not part of the SWS but could have been potentially sampled for the SWS because they were part of the SWS frame.[8]

In Rounds 1 through 4, we used data from SSA on the counts of eligible beneficiaries in each county in 2003 to form 1,330 PSUs, each of which consisted of one or more counties. In 2012, prior to Round 5, we studied the distribution of SSI and SSDI beneficiaries in the 2003 PSUs using 2011 data and found that, although the total numbers had changed from 2003 to

[5] Only SWS members who were working at the time of the Round 6 interview are eligible for the longitudinal sample in Round 7. A new cross-sectional SWS sample will also be included in the Round 7 SWS.

[6] The RBS and the main sample of the SWS involved selecting individuals within selected clusters of geographic areas, and is therefore referred to as a "clustered sample." The supplemental sample (for the SWS only) was selected across the entire population of successful workers and was therefore not limited to those residing in selected clusters. It is therefore referred to as an "unclustered sample." This is discussed in detail later.

[7] The original selected samples for the RBS and SWS, and the frames from which they were selected, improperly included a very small number of cases whose ineligibility was known prior to sample selection. All sample and frame counts in this report exclude these cases.

[8] There were an additional 21 sampled cases in the RBS, of which 7 responded, that were part of the SWS frame, but were not sampled for the SWS.

2011, the distributions did not change very much. Therefore, we selected a new sample of PSUs in Round 5 from the same group of 1,330 PSUs that were formed in prior to Round 1 (in 2003). As stated earlier, we used the same PSUs in Round 6 (for both the RBS and the SWS main sample) that we had selected in Round 5.

For the RBS in Round 6, we fielded a nationally representative sample of 7,947 SSA disability beneficiaries. Except for the way we stratified the sample of the PSUs,[9] the sample design for the RBS was nearly identical to the design of the RBS in Rounds 1 through 5. 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, 2016.[10] 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) age 18 to 29, (2) age 30 to 39, (3) age 40 to 49, and (4) age 50 and older. To ensure a sufficient number of persons seeking work, we oversampled beneficiaries in the first three cohorts (age 18 to 49). The target number of completed interviews for Round 6 was 1,111 beneficiaries in each of the three younger age groups. For those age 50 and older, the target number of completed interviews was 667 beneficiaries. We summarize the actual sample sizes and number of completed interviews for both the RBS and SWS under the revised Round 6 design in Table I.2.

The SWS was limited to SSI and SSDI beneficiaries who were eligible for the RBS, but were considered “successful workers” because their earnings for a sustained period were sufficiently high. In particular, the SSI and SSDI beneficiaries were required to (1) have earnings above SSA’s non-blind substantial gainful activity (SGA) monthly earnings level ($1,130 in 2016 and $1,170 in 2017) for a minimum of three consecutive calendar months at any time between August 1, 2016 and July 31, 2017, and (2) be younger than age 62 on June 30, 2016. The successful work must have occurred within a time frame so that the vast majority would be interviewed within six months of the end of their successful work (if they were not currently working), and their earnings had to have been revealed in the Disability Control File (DCF) at the time of data extraction—removing from the population any successful workers who had a long delay in having their earnings recorded on the DCF.[11] To ensure a large enough number of

[9] As noted earlier, the sample design for Rounds 1 through 4 included two samples, one for all beneficiaries (the RBS) and one for the participants in the TTW program (the Ticket Participant Sample). To accommodate the rollout of the TTW program, the PSUs were sampled within strata defined by the three phases of the rollout. The sample design for Round 5 only included one sample, that of all beneficiaries. The PSUs were not drawn within strata, except those defined by the two certainty PSUs. The Round 6 sample used the same PSUs as those sampled in Round 5.

[10] 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.

[11] Some SSI and SSDI beneficiaries would be considered successful workers because their earnings and age met the threshold, but they had to be excluded from the sample frame for the sampling effort due to a delay in recording their earnings on the DCF. For these individuals, a lag of up to six years would exist between the time that they received their earnings, and the time that the earnings data were recorded in the DCF, though most had their earnings recorded after three years. There was no way they could be identified in time for the data extraction. In November 2020, we conducted an updated extract of DCF earnings data for the time period in question, and post-stratified the analysis weights to these new totals.

successful workers for sampling, we formed seven successive frames of successful workers over time. Each one was revealed by comparing the full sampling frame to updated earnings information and identifying all successful workers at that time, then removing them from subsequent frames to make the frames mutually exclusive. The SWS sampling frames were all subsets of the same sampling frame used for the Round 6 RBS sample, and are therefore referred to as "extracts" from the larger frame. Using these constraints to define the target population, we identified a population of 89,936 successful workers.[12] Within each of the seven extracts, we stratified the SWS into two strata defined by beneficiary type (SSDI only, and SSI, which included both SSI only and concurrent beneficiaries) and selected a probability sample from each extract. From these extracts, we fielded a nationally representative sample of 13,271 successful workers.[13] We included one screening question as an additional constraint: the sampled successful workers had to indicate that they had been working in the past six months.[14] The targeted number of completed interviews for the two strata was 2,250 interviews apiece across all extracts. We did not know the size of each extract before sample selection; the first sample size allocation to the samples in each extract was based on historical data. After the release of each extract, the allocation of sample sizes to the samples from the remaining extracts was adjusted to make the allocation as proportional as possible to the population of successful workers over time, within each of the two beneficiary type strata (SSDI only and SSI). We did not complete sample selection until after the release of the last extract.

Because of the concerns about the number of successful workers within strata and their distribution across PSUs within each extract, we decided to supplement the main SWS (within the PSUs) with a second independent sample of successful workers. This supplemental sample was divided into two geographic strata (successful workers residing in a sampled PSU, and successful workers not residing in any of the sampled PSUs).[15] We refer to the multistage sample design as the "clustered" sample, and to the second independent sample as the "unclustered"

[12] This total did not include successful workers whose earnings were not included in the DCF at the time of extraction due to a lag in the posting of earnings for some beneficiaries. Furthermore, it did include a small number of cases (4,746 out of 89,936) that met the successful work criteria at the time of the initial extraction, but in an updated extraction from November 2020, were found to not meet the criteria during the time period in question. In the later extraction, the actual weighted total number of successful workers was found to be 288,576. We post-stratified the provisional analysis weights to match this total.

[13] For reasons explained later in this chapter, this sample includes 490 duplicates. As a result, 12,781 unique cases were sampled. As noted in Section I.A.2, the frame from which the SWS was drawn was provisional. In an updated extraction from November 2020, we found that 725 of the 13,271 sample cases, including 219 of the 4,587 completed interviews, did not meet the criteria for successful work. In the updated final analysis weights, where the provisional analysis weights were post-stratified to totals from the November 2020 frame, these 725 sample members were given zero weight.

[14] This screening question was included to account for situations where a long period of time had elapsed between the date when the case was released for data collection and the interview date. Few cases were actually removed from the sample due to this screening question, especially in later extracts.

[15] Given that the target population for the NBS did not include Puerto Rico or other outlying territories, we excluded from the frame all beneficiaries and successful workers who resided in these areas.

sample.[16] We call the combination of data from the clustered and unclustered samples to calculate estimates a “dual” sample design. The clustered sample included in-person follow-up for sample members who could not be located or otherwise did not respond by phone; the unclustered sample did not have in-person follow-up.

After the completion of the sample selection for all seven extracts, we created a single set of SWS composite weights that combined information from the clustered and unclustered SWS, which appropriately accounted for the different follow-up rules between the two samples.[17] Table I.2 includes the total across the two samples in the SWS, and does not break out the counts between clustered and unclustered samples; the 490 duplicate cases that were selected for both the clustered and unclustered samples are counted twice in this table. The dual sample design and the calculation of the composite weights that combine the weights from the clustered and unclustered sample are discussed in detail in Chapter III, and the counts within the clustered and unclustered sample are also provided in Chapter III.

C. Round 6 survey overview

The NBS was designed and implemented to maximize both response and data quality. Table I.3 describes the most significant sources of potential error identified at the outset of the NBS and how we attempted to minimize the impact of them. A more detailed discussion of our approach to minimizing total survey error can be found in Appendix A of the Round 6 User’s Guide (Callahan et al. 2019).

[16] Because of the small populations of successful workers, Mathematica often selected successful workers who resided in both the selected PSUs for the clustered and in-PSU strata of the unclustered samples. Hence, we had to account for these duplicate cases in the weighting process (discussed later).

[17] These composite weights, combining weights from the clustered and unclustered samples in the SWS, should not be confused with the composite weights that combined the RBS sampling weights and the SWS sampling weights that we briefly alluded to in the introductory paragraphs.

Table I.2. NBS–General Waves (RBS and SWS) Round 6 actual sample sizes, target completes, and completes

Sampling strata	Selected sample size[a]	Original target completed interviews[b]	Actual completed interviews[c]
Representative beneficiary sample	7,947	4,000	4,002
18- to 29-year-olds	2,356	1,111	1,120
30- to 39-year-olds	2,243	1,111	1,081
40- to 49-year-olds	2,153	1,111	1,129
50-year-olds or older	1,195	667	672
Successful worker sample			
December 2016 extract	2,647	631	982
SSDI only	1,123	250	397
SSI (SSI only + concurrent)	1,524	381	585
January 2017 extract	2,095	737	723
SSDI only	1,017	344	336
SSI (SSI only + concurrent)	1,078	393	387
March 2017 extract	1,890	773	740
SSDI only	873	373	351
SSI (SSI only + concurrent)	1,017	400	389
April 2017 extract	1,607	627	606
SSDI only	854	344	324
SSI (SSI only + concurrent)	753	283	282
June 2017 extract	1,849	657	582
SSDI only	922	350	313
SSI (SSI only + concurrent)	927	307	289
July 2017 extract	1,373	573	442
SSDI only	895	315	283
SSI (SSI only + concurrent)	478	258	159
September 2017 extract	1,807	502	512
SSDI only	1,123	274	324
SSI (SSI only + concurrent)	684	228	188
Total	13,271	4,500	4,587
SSDI only	6,807	2,250	2,328
SSI (SSI only + concurrent)	6,464	2,250	2,259

Source: NBS Round 6 (the second round of NBS–General Waves).

[a] The 13,271 SWS sample cases include 725 that were later found to not be successful workers

[b] The target completed interviews for the SWS shown here were calculated prior to receiving the first extract, using historical data from simulated successful worker populations in 2011-12, 2013-14, and 2015-16. In fact, there were actually seven allocations, with a new sample allocation calculated after the population sizes for each extract were revealed. This explains the sometimes large deviation between the target allocation and the actual number of completed interviews.

[c] The 4,587 SWS completed interviews include 219 that were later found to not be successful workers. In the final post-stratification, these cases had zero weight.

Table I.3. Sources and descriptions of potential error 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 (PSUs) 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 that are 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
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 is 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.	Use of same instrument in both interview modes; use of probes; adaptive equipment; interviewer training, validation of field interviews; assistance during interview; use of proxy, if sample member is 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 selected items asking for wages and household income, as well as cohabitation status.[18] Unit nonresponse was the greater concern given the population; thus, the survey was designed with a dual-mode approach. Mathematica made all initial attempts to interview beneficiaries using CATI. We sought a proxy respondent when a sample member was unable to participate in the survey

[18] Round 6 is the first round where a high level of missingness in C_Cohab (the cohabitation status variable) was observed. It was due to an error in the Round 6 questionnaire skip logic. Details about this error are provided in Chapter IV.

because of his or her disability. To promote response among Hispanic sample members whose primary language is Spanish, Mathematica provided the questionnaire in Spanish. For languages other than English or Spanish, interpreters, if available in the sample member's home, helped to conduct the 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. After locating the sample member, the field locator attempted to facilitate an interview with him or her 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 both modes.[19]

We began Round 6 CATI data collection for the NBS in February 2017. In April 2017, Mathematica began in-person locating and CAPI, which continued concurrently with CATI through November 2017.

1. Completes and Response Rates

In total, Mathematica completed 8,589 interviews across the RBS and SWS (including, 131 partially completed interviews)—4,002 from the RBS and 4,587 from the SWS.[20] An additional 290 beneficiaries from the RBS and 463 successful workers were deemed ineligible for the survey.[21] Because of the independence of the RBS and SWS sample selections and the independence of the clustered and unclustered sample selections within the SWS, individuals could be selected for more than one sample. Therefore, the number of unique completed interviews was 8,410.[22] Across both samples, Mathematica completed 8,402 cases by CATI

[19] No CAPI interviews were conducted for unclustered sample cases in the SWS unless an unclustered sample case was duplicated with a clustered sample case. In that event, the information from the field effort was not used for the unclustered case.

[20] As noted in Section I.A.2 and I.B, the frame from which the SWS was drawn was provisional. In an updated extraction from November 2020, we found that 725 of the 13,271 sample cases, including 219 of the 4,587 completed interviews, did not meet the criteria for successful work. In the updated final analysis weights, where the provisional analysis weights were post-stratified to totals from the November 2020 frame, these 725 sample members were given zero weight.

[21] Ineligible sample members include those who were deceased, incarcerated, in active military, or no longer living in the continental United States and those whose benefit status was pending at the time of the interview. For the SWS, ineligibles also included sample members who had not worked in the past six months at the time of the interview.

[22] Among sample cases that were completed interviews only, there were 38 duplicates between the RBS and SWS (76 sample cases total) and 141 duplicates (282 sample cases total) between the clustered and unclustered samples within the SWS, for a total of 179 duplicates. The counts of ineligible cases included 15 duplicates; the number of unique ineligible cases across both samples was 738.

(either directly from the survey operations center or via field staff who handed respondents a cell phone) and 8 cases by CAPI.[23]

The unweighted and weighted response rates for the RBS were 54.0 and 58.8 percent, respectively. For the SWS, the unweighted and weighted response rates were 38.0 and 41.3 percent, respectively.[24]

2. Nonresponse bias

Because the weighted response rates were less than 80 percent for both samples, we conducted a nonresponse bias analysis at the end of data collection. We examined all 7,947 selected sample cases in the RBS and all 13,271 selected sample cases in the SWS to determine if there were systematic differences between respondents and nonrespondents for a variety of covariates.[25] Our analysis revealed differences between respondents and nonrespondents for some variables, but the nonresponse adjustments to the weights appear to have eliminated all such differences in both samples.

There were other potential sources of bias for some small populations representing county-level economic indicators, but this was unrelated to nonresponse. In these cases, the weighted estimates of the small populations differed from those in the frame because we did not control for those populations when we created the initial sampling weights. This was because the variables representing these populations (1) were not considered important enough to be used in post-stratification, relative to the variables we used for this adjustment, and (2) were not included as covariates in the final nonresponse models, generally because the samples were too small. We therefore could not reconcile these differences when adjusting these weights for nonresponse or when post-stratifying them to marginal population totals.

The full nonresponse bias analysis can be obtained from SSA (https://www.ssa.gov/disabilityresearch/nbs_round_6.html).

[23] We reserved CAPI mode for special situations in which respondents were unable to complete the interview by using another method; only eight respondents requested an in-person interview. Of the 8,402 CATI completes, 1,396 were call-ins from the field that were a direct result of field locating, while another 532 were sent to the field at some point.

[24] Details about the formulas used to calculate the response rates, and alternative formulas that could have been used, are given in Chapter III. Using information from the updated frame from November 2020, the updated weighted SWS response rate was 40.8 percent. This reduction of 0.5 percent was due to the fact that a large percentage of the 725 sampled cases who were not successful workers were found to be ineligible at data collection. Removing these sample cases had a negative effect on the weighted response rate.

[25] The nonresponse bias analysis was conducted on all 13,271 sample cases. However, in the updated extraction from November 2020, 725 sampled cases did not meet the criteria for successful work. In the final post-stratification, the weights for these cases were set to zero.

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 editing and coding conventions that were used in the data files.

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 conducting a meticulous review of individual responses, we acknowledge that some suspect values remain on the file. (See McDonald et al. [2019] 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 “don’t know” (.D) 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 for the same conditions 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, Item B29 asks, "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. For example, Item G53 asks "Thinking only about the services {you/NAME} used in 2016, what are the main reasons {you/he/she} decided to use these services?" Interviewers then coded the verbatim response into seven established categories. If the response did not fit into one of the categories, interviewers selected "other."

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. Coders used the Ascribe coding software to apply codes to verbatim responses. The Ascribe program allowed coders to sort and filter verbatim responses in several ways to facilitate the coding effort. We sorted verbatim responses alphabetically by item for coders. Records could also be filtered to show responses that had been reviewed by a supervisor, or to show cases with clarifying notes for a coder. . 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 McDonald et al. [2019] 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 Round 6 (the second round of NBS–General Waves).

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 he or she 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 constructed variables for 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).

In prior rounds of data collection, we coded respondents' verbatim responses by using the International Classification of Diseases, 9th Revision, Clinical Modification (ICD-9) five-digit coding scheme. 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. A newer version of the coding scheme (ICD-10) was released prior to Round 6 of data collection. Rather than switching to the ICD-10, which included a new layout of the codes and more complex mapping, SSA agreed that we should use a broader, three-digit coding scheme derived from the ICD-9 categories for Round 6. The list of 21 codes used for Round 6 of data collection is included in Table II.2. The coders, many of whom had medical coding experience, attended a four-hour training session before they started coding; they also attended weekly check-in meetings with coding supervisors throughout the coding effort. 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). Each code was applied a maximum of one time per question, even in instances where the same medical code could be applied to more than one condition reported within a question. For instance, "bipolar" and "schizophrenia" are distinct conditions that fall under the same medical code (050 – mental disorders). If both conditions were reported within the same response, "bipolar" and "schizophrenia" would receive code 050 one time. If each condition was reported in a separate question (for instance, if the respondent reported "bipolar" at Item B2 and "schizophrenia" at Item B4), both conditions were coded.

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. 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.

Table II.2. Round 6 Coding Scheme

Code	Label	Description of ICD-9 Codes	Corresponding ICD-9 Codes
010	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
020	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
030	Endocrine/nutritional disorders	Thyroid disorders, diabetes, abnormal growth disorders, nutritional disorders, and other metabolic and immune disorders	240.0–279.9
040	Blood/blood-forming diseases	Diseases of blood cells and spleen	280.0–289.9
050	Mental disorders	Psychoses, neurotic and personality disorders, and other non-psychotic mental disorders. EXCLUDES Intellectual disability (formerly termed mental retardation)	290.0–302.9, 305.00–314.9, 315–316
051	Intellectual disability	Intellectual disability	317.0-319.9
060	Diseases of nervous system	Disorders of brain, spinal cord, central nervous system, peripheral nervous system, and senses, including paralytic syndromes	320.0–359.9
061	Diseases and disorders of the eye and ear	Disorders of eye and ear	360.0–389.9
070	Diseases of circulatory system	Heart disease; disorders of circulation; and diseases of arteries, veins, and capillaries	390-459.9
080	Diseases of respiratory system	Disorders of the nasal, sinus, upper respiratory tract, and lungs, including chronic obstructive pulmonary disease	460-519.9
090	Diseases of digestive system	Diseases of the oral cavity, stomach, esophagus, and duodenum	520.0-579.9
100	Diseases of genitourinary system	Diseases of the kidneys, urinary system, genital organs, and breasts	580.0-629.9
110	Complications of pregnancy, child birth, and puerperium	Complications related to pregnancy or delivery and complications of puerperium	630-677
120	Diseases of skin/ subcutaneous tissue	Infections of the skin, inflammatory conditions, and other skin diseases	680.0-709.9
130	Diseases of musculoskeletal system	Muscle, bone, and joint problems, including arthropathies, rheumatism, osteopathies, and acquired musculoskeletal deformities	710-719, 725-739
131	Diseases of the musculoskeletal system: back disorders.	intervertebral disc disorders, other disorders of cervical region, and other and unspecified disorders of the back	720-724
140	Congenital anomalies	Problems arising from abnormal fetal development, including birth defects and genetic abnormalities	740.0-759.9

Code	Label	Description of ICD-9 Codes	Corresponding ICD-9 Codes
150	Conditions in the perinatal period	Conditions that have origins in birth period, even if disorder emerges later	760.0-779.9
160	Symptoms, signs, and ill-defined conditions	Ill-defined conditions and symptoms; used when no more specific diagnosis can be made	780.01-799.9
170	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
180	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 Round 6 (the second round of NBS–General Waves).

3. Industry and Occupation

In Section C of the questionnaire, we collected information about a sample member’s current employment. In Section C_B of the questionnaire, we collected information about a sample member’s employment in the last 6 months, if the sample member was not currently working at the time of the interview. In Section D of the questionnaire, we collected information about a sample member’s employment in 2016. For each job, respondents were asked to report their occupation (Items C2, C_B2, and D4) and the type of business or industry (Items C3, C_B3, and D5) in which they were employed. In previous rounds of data collection, we used the Bureau of Labor Statistics 2000 Standard Occupational Classification (SOC) to code verbatim responses to these items. For Round 6, we used the Bureau of Labor Statistics 2010 Standard Occupational Classification (SOC) for coding.[26] 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). Round 6 codes applied using the 2010 SOC remain comparable with earlier rounds coded using the 2000 SOC, as all major and minor group

[26] For more information, see *Standard Occupational Classification Manual, 2010,* or http://www.bls.gov/soc.

codes remained consistent across both coding schemes. We list the three-digit minor groups that are classified within major groups in Appendix B.

In previous rounds of data collection, we coded verbatim responses to the industry items according to the 2002 North American Industry Classification System (NAICS). For Round 6, we used the 2017 North American Industry Classification System (NAICS).[27] 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. Round 6 codes applied using the 2017 NAICS remain comparable with earlier rounds that used the 2002 NAICS, as all industry sector and subsector codes remained consistent across both coding schemes. (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.

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.

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.

[27] For more information, see North American Industry Classification System, 2017, or https://www.census.gov/eos/www/naics/index.html

III. SAMPLING WEIGHTS

We determined the final analysis weights for the Representative Beneficiary Sample (RBS) and Successful Worker Sample (SWS) 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 Sections B and C, respectively, we describe the procedures for computing the weights for the RBS and SWS in more detail.

A. Computing and adjusting the sampling weights: A summary

1. Representative Beneficiary Sample

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.[28] When preparing for Round 1 in 2003, we formed 1,330 PSUs, each of which consisted of one or more counties, by using data from SSA on the counts of eligible beneficiaries in each county. For Rounds 1 through 4, we selected PSUs only once (in 2003) from this list of PSUs. When preparing for Round 5 of the NBS–General Waves in 2014, the first-stage sampling units were selected from the same list of PSUs.[29] These PSUs from Round 5 were used as the first-stage sampling units for Round 6, and will be the first-stage sampling units for Round 7. 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.[30] 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 many more beneficiaries than other counties. Therefore, we partitioned them into a large number of secondary sampling units (SSUs) based on beneficiary zip codes.[31] From these SSUs, we

[28] In two 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.

[29] 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 Rounds 1 through 4. Although the set of PSUs from which to sample did not change from Rounds 1 through 4 to Round 5, 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.

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

[31] 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 prior to Round 1. But we selected a new set of SSUs for the Round 5 sample by using a measure of size for each SSU that was based on the most current counts of beneficiaries, and used those same selected SSUs for Round 6.

selected four SSUs from the Los Angeles County PSU and two from the Cook County PSU.[32] 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, 2016.[33] We used four age-based strata in each PSU. In particular, we stratified beneficiaries into the following age groups: (1) age 18 to 29, (2) age 30 to 39, (3) age 40 to 49, and (4) age 50 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.[34]

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.[35] 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 adjusted weight is the product of the base weight and the adjustment factor. One would form the

[32] 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.

[33] 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, or 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.

[34] 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).

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

"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 by: (1) estimating a propensity score for locating a sample member and (2) 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 their phone number or may frequently move from one address to another, but there is no evidence to suggest that—once located—they 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 2016-2017), 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 (using the STEPWISE option of the SAS LOGISTIC procedure with weights normalized to the sample size), 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 procedures in SAS do not fully account for the complex survey design, we developed the final weighted models by using software that does account for the complex sample design (the RLOGIST procedure in SUDAAN and the SURVEYLOGISTIC procedure in SAS).

The next step called for carefully evaluating a series of models by comparing the following measures of predictive ability and goodness of fit: the R-squared statistic, the percentage of concordant and discordant pairs, and the Hosmer-Lemeshow (H-L) goodness-of-fit test.[36]

[36] In Rounds 1 through 5, we also used Akaike's Information Criterion, or AIC, as a model diagnostic (discussed in Akaike 1974). We obtained the AIC from SAS output of the LOGISTIC procedure, since it is not available in

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.[37] We then used the specific covariate values for each located person to estimate the propensity score, and used the inverse of the propensity score to determine the adjustment factor. 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 for 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 1.07.

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,[38] and five categories of annual earnings from the Disability Control Files (DCF) of 2015 and 2016.[39] After post-stratification, we checked the survey weights again to determine

SUDAAN. However, in Round 6, we began using the SURVEYLOGISTIC procedure in SAS, which does account for the survey design, and the AIC in these procedures was not helpful as a model diagnostic.

[37] 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 avoiding the first two was the 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.

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

[39] This was an attempt to address small negative bias in annual earnings, which was observed in Rounds 1 through 4. We arrived at the five earnings categories used in Round 5 after a lengthy investigation using both (annual) IRS and (monthly) DCF earnings. Using data from the 2014 sampling frame, we calculated the percentage 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 post-stratification 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, which raises timing issues, and dilutes 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 cut points given above because using them resulted in estimated annual earnings that were closest to the IRS values. The 2013 data were used because of a lag in identifying earnings in the 2014 data, which did not

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

2. Successful Worker Sample

We defined successful workers in the introduction as Supplemental Security Income (SSI) or Social Security Disability Insurance (SSDI) beneficiaries who were (1) active or in suspense on June 30, 2016, (2) with earnings above SSA's non-blind substantial gainful activity (SGA) [40] earnings level for a minimum of three consecutive calendar months at any time between August 1, 2016 and July 31, 2017, and (3) were less than 62 years old on June 30, 2016. The earnings for each successful worker had to have been revealed in the DCF at the time of data extraction—removing from the population eligible for sampling in that extract any successful workers who had a long delay in having their earnings recorded on the DCF. These successful workers were accounted for in a subsequent extraction, in November 2020. The (provisional) analysis weights for sampled cases were post-stratified again to match the total number of successful workers in that later extract. Finally, for each extract, we needed to ensure that the potential elapsed time period between the final identified month of the successful work period and the interview date did not exceed six months (in most cases).[41] This means that each extract had to be limited to successful workers whose successful work ended late enough to satisfy this requirement. The data for each successive frame were extracted at (approximately) six week intervals, to ensure that enough new successful workers could be identified in each new extract. For the first six of the successive frames, data were extracted on the Monday or Tuesday after the following dates: December 1, 2016, January 15, 2017, March 1, 2017, April 15, 2017, June 1, 2017, and July 15, 2017. Due to the short data collection window available for successful workers in the final extract, we performed the extraction for the final frame on the Tuesday before September 1, 2017 (August 29). Table III.1 summarizes the earliest acceptable final month of successful work for a successful worker to be included in each extract. Also included in this table is the first month of ineligibility for those whose successful work actually ended on the earliest acceptable final month shown. For those who met these criteria to be included in the extract, sample members were asked in the questionnaire if they had worked in the past six months. If they answered negatively, they were screened out.

have complete information on the amount of earnings that beneficiaries received in that year. For the Round 6, we determined five earnings categories using earnings data from the 2015 and 2016 DCF files.

[40] This threshold was $1,090 in 2015 and $1,130 in 2016.

[41] As per SSA's specifications, the period between the last month of successful work and the interview date was limited to six months to avoid issues of recall about the sample member's successful work period. We say "in most cases" because it was possible, though unlikely, for the sample member from the first few extracts to have had their successful work cease more than six months ago. For this to occur, (1) the interview had to occur long after the case was released for data collection, meaning that this was only possible in one of the earlier extracts, (2) their successful work didn't continue, but ceased long before data collection, and (3) they did not answer the screening question correctly about whether they worked in the past six months, or their work in the past six months did not exceed the SGA threshold.

Table III.1. Earliest acceptable final identified month of successful work for each extract, and resulting first month of ineligibility

Extract	Earliest acceptable final month of successful work	First month of ineligibility for those with earliest acceptable final month of successful work
December 1, 2016	October, 2016	May, 2017
January 15, 2017	November, 2016	June, 2017
March 1, 2017	December, 2016	July, 2017
April 15, 2017	February, 2017	September, 2017
June 1, 2017	March, 2017	October, 2017
July 15, 2017	May, 2017	December, 2017[a]
September 1, 2017	June, 2017	January, 2018[a]

[a] The first month of ineligibility for the July and September extracts occurs after the end of the data collection period.

The window of time that a successful worker could be identified for inclusion in an extract, selected for the sample, and have an attempted interview, is illustrated in Figure III.1 for three of the seven extracts. The figure shows the length of time between the successful work and the interview, and how this elapsed time must not exceed six months. The first oval corresponds to the first sample extract, which is limited to those whose successful work either ended in October or November in 2016, or continued at the time of the extract creation in early December. It excludes those whose three consecutive months of successful work ended earlier than October, 2016. This is because, for the December extract, we estimated that the successful workers' interview date could be as late as April 2017. For someone whose successful work ended in September, this would be more than six months of recall. It is possible that the interview date would be sooner than April 2017, in which case we would be excluding someone from the frame whose successful work ended fewer than six months beforehand. By the same token, if the interview was in May, someone whose successful work ended on October 31 would have more than a six-month gap until the interview date (and would be screened out from the screener question in the questionnaire). However, constructing the frames in this way ensures that most will have a gap that is less than six months, and that few cases would be screened out based on the response to the screening question in the questionnaire.

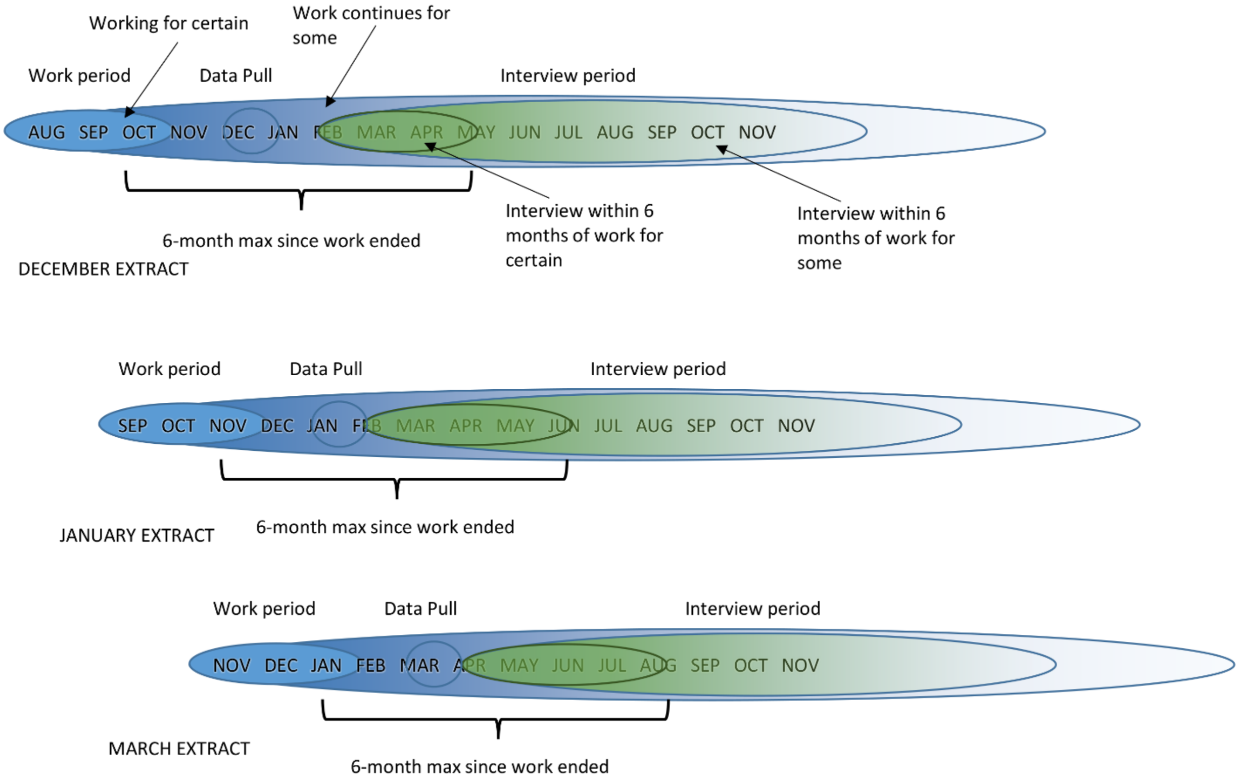
As with the RBS, we used the PSUs as the primary source of sample members for the SWS and selected an initially larger (augmented) sample. We selected the sample of successful workers from among the identified successful workers residing in the same PSUs that were selected for the RBS, and used no SSUs.[42] Within each of the seven extracts, we stratified the SWS into two strata defined by beneficiary type (SSDI only, and SSI, which included both SSI only and concurrent beneficiaries).

Because of concerns about the small numbers of successful workers within each stratum and their distributions across PSUs within each extract, we decided to supplement the sample within the PSUs with a second independent sample of successful workers from two geographic strata

[42] For the SWS, Mathematica selected successful workers from the entire Los Angeles County PSU and from the entire Cook County PSU. In the RBS we subsampled SSUs in these two counties.

defined by the PSUs (successful workers residing in a PSU or not residing in any of the PSUs).[43] We refer to the initial sample design as the "clustered" sample; the second independent sample is referred to as the "unclustered" sample.[44] The clustered sample therefore had two strata within each extract (SSDI only and SSI), and the unclustered sample had four strata (the cross-classification of the SSDI/SSI variable and the geographic location variable). We refer to the combination of data from the clustered and unclustered samples to calculate estimates as a dual sample design (discussed later).

Figure III.1. Timeline for extracts in Successful Worker Sample, including work period, data pull dates, and admissible data collection period for each extract


Working for certain
Work continues for some
Work period
Data Pull
Interview period
AUG SEP OCT NOV DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV
6-month max since work ended
Interview within 6 months of work for certain
Interview within 6 months of work for some
DECEMBER EXTRACT
Work period
Data Pull
Interview period
SEP OCT NOV DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV
JANUARY EXTRACT
6-month max since work ended
Work period
Data Pull
Interview period
NOV DEC JAN FEB MAR APR MAY JUN JUL AUG SEP OCT NOV
MARCH EXTRACT
6-month max since work ended

Note: Solid ovals identify the "for certain" periods, and gradients represent the decline in certainty over time.

We computed the initial sampling weights for the SWS on the basis of the inverse of the selection probability for the successful worker within each extract. As with the RBS, we computed the weights for the augmented sample and then adjusted them for the number of sample members released into the final sample. (In the case of the SWS, we did not release any additional sample cases after the initial release for each extract.) We adjusted for located sample

[43] Given that the target population for the NBS did not include Puerto Rico or other outlying territories, we excluded from the frame all beneficiaries and successful workers who resided in these areas.

[44] Because of the small populations of successful workers, Mathematica often selected successful workers who resided in both the selected PSUs for the clustered and in-PSU strata of the unclustered samples. Hence, we had to account for these duplicate cases in the weighting process (discussed later).

members and then for response among such members. We used logistic propensity models to calculate the location adjustment for all successful workers and the response adjustments for located successful workers. The modeling procedures were similar to those used with the RBS, discussed in Section A.1 of this chapter.

For the sake of efficiency, we combined the seven extract samples into a single sample when calculating the nonresponse adjustments. Within each stratum, we trimmed the weights to ensure that the design effect was not adversely affected by outlier weights. (In Section C, we provide more detail on the trimming of successful workers' weights and the design effects attributable to unequal weighting before and after trimming.) We also conducted a single post-stratification across the seven extract samples. In this process, we adjusted the weights so that the marginal totals matched the frame totals within subgroups defined by five earnings categories,[45] the four age categories, program title,[46] and the extract totals. After post-stratification, we checked the survey again to determine the need for more trimming. Even though the Round 6 weights required trimming before post-stratification in the SWS, they required no further trimming after post-stratification. Much later, in November 2020, we conducted a final extraction from the DCF, and post-stratified the weights (again) to new marginal totals within subgroups defined by the same five earnings categories (with updated values), four age categories, program title, five disability categories,[47] and gender. This final post-stratification is described in detail in Section III.C.5.

To calculate the weights for the SWS, it was necessary for us to create composite weights that combined the sampling weights from the clustered and unclustered components.[48] The procedure for calculating the SWS composite weights is discussed later in this chapter.

3. Composite weights for combining samples (SWS and RBS)

Although the successful worker population constitutes a small subset of the beneficiary population, some analyses required a sample with a substantial number of individuals both within and outside the successful worker population. Such a sample simply represents a combination of the successful worker and beneficiary samples and required the use of another type of composite weights to account for the combined sample. When conducting analyses representing the beneficiary population, we used the combined sample weights to make estimates comparing successful workers to others within the beneficiary population. (Analyses limited to the successful workers' subpopulation used weights from the SWS only.)

[45] The five earnings categories used for post-stratification in the SWS differed from those used for the RBS. In the RBS, most sample members did not have earnings. However, by definition, nearly everyone in the SWS had earnings in 2015 and 2016, so the categories were reconfigured to accommodate this.

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

[47] The five disability categories were the same as those used in the nonresponse adjustments.

[48] This is referring to the creation of weights that combine the unclustered and clustered samples from the SWS. In the next section, we discuss the creation of composite weights that are used to combine the weights from the RBS and SWS. These two sets of composite weights are distinct and should not be confused.

In Round 1, some analyses required a combination of data from the RBS and TPS, similar to the RBS-SWS combined sample described above. To create the composite weights for that combined sample, we used a sophisticated procedure—similar to that used to combine the clustered and unclustered samples in the SWS—in order to minimize the variance of survey estimates. The procedure allowed weights to be applied to observations duplicated across the two samples.[49] However, given that the Ticket participants were such a small fraction of the beneficiary sample frame, we used a simpler alternative method in Rounds 2 through 4.

In Round 6, we used this simpler alternative again when creating RBS-SWS composite weights. We replaced the original RBS weights with a value of zero among the 112 sample members who happened to be successful workers but were not necessarily sampled in the SWS.[50] To ensure representation of the successful worker population, these 112 members of the RBS were represented by the 4,587 members of the SWS who had completed an interview (or had ineligible dispositions after sample selection). The sum of the weights for the 112 successful workers in the RBS is an unbiased estimate of the number of successful workers in the sampling frame. However, given the relatively small number of successful workers in the RBS, the estimate did not equal the known total in the sampling frame. The post-stratification adjustment realigned the population totals so that the weights for the responding cases in the SWS added up to the total SWS population, and the weights for the non-SWS cases in the RBS added up to the total non-SWS population. In November 2020, we re-created the RBS-SWS composite weights using the same procedures with the new SWS sample frame and new SWS analysis weights.

4. 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. Computing weights for the Representative Beneficiary Sample

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,[51] 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 youngest age strata (18- to 29-year-olds) was 3,385 sample members, and for the two middle age strata (30- to 39-year-olds and 40- to 49-year-olds) the sample sizes were 3,272 and 3,278

[49] A complex procedure also combined the clustered and unclustered samples of the SWS (described in Section C of this chapter).

[50] Of the 112 successful workers in the RBS, 91 were also part of the SWS.

[51] 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.

sample members, respectively. The average across these three age groups was roughly three times the target sample size of 1,111. For beneficiaries age 50 and older, the augmented sample size was 1,991 (again, about three 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 11,926 sample members by taking the inverse of the augmented sampling rate (Fj) for each stratum. In Table III.2, we provide the augmented sampling rates and initial weights, as well as the sizes of the population, augmented sample, and released sample.

Table III.2. Study population (as of June 30, 2016), initial augmented sample sizes, and initial weights by sampling strata in the National Beneficiary Survey

Sampling strata (ages as of June 30, 2016)	Study population	Augmented sample size	Augmented sampling rate (*Fj*)	Initial sample weights	Released sample
Beneficiaries age 18 to 29	1,382,706	3,385	0.002449	408.48	2,356
Beneficiaries age 30 to 39	1,470,933	3,272	0.002224	449.55	2,243
Beneficiaries age 40 to 49	2,201,196	3,278	0.001489	671.51	2,153
Beneficiaries age 50 to FRA	8,784,221	1,991	0.000227	4412.0	1,195
Total	**13,839,056**	**11,926**			**7,947**

Source: Study population counts are from SSA administrative CERs and DBADs files, extracted for NBS Round 6. SSA determined the number of complete interviews based upon recommendations from Mathematica.

FRA = full retirement age.

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 7,947 beneficiaries, as shown in Table III.2.

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 people 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 were in the SSA files for the beneficiary; (5) the living situation of the beneficiary; (6) the program(s) through which the beneficiary received benefits (SSI, SSDI, or both); (7) primary disability, and (8) 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 Sample 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 who were 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)[52]
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.

[52] A located passive refusal is a case where we made contact with the sample member or a gatekeeper associated with the sample member, but the case passively refused by not responding to later outreach attempts.

b. Response Rates

The 58.8 percent response rate for the RBS (Table III.3 is the weighted[53] count of sample members who completed an interview or were deemed ineligible divided by the weighted sample count of all sample members.[54] It can be approximated by taking the product of the weighted location rate and the weighted cooperation rate among located sample members.[55]

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 94 percent (Table III.3). The weighted cooperation rate (that is, the weighted cooperation rate among located sample members) of 63 percent (Table III.3) 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.[56] Weighted cooperation rates reflect the rate at which completed interviews are obtained from repeated contact efforts among located persons.

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

[54] 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 very close in value to the American Association of Public Opinion Research (AAPOR) standard response rate calculation: 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) incorporating the ineligibles into the numerator and denominator of the response rate is equivalent to the definition of a more conventional response rate, when all nonrespondents have unknown eligibility status. In our case, the vast majority of nonrespondents have unknown eligibility status.

[55] 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.

[56] The counts provided in Table III.3 are unweighted, and the rates (percentages) are weighted by the original sampling weight for the location rate, and the location-adjusted weight for the cooperation rate. The final response rate is weighted using the original sampling weight.

Table III.3. Weighted location, cooperation, and response rates for Representative Beneficiary Sample, by selected characteristics

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Weighted location rate	Count	Weighted cooperation rate	Weighted Response rate
All	7,947	7,332	94.1	4,292	62.5	58.8
SSI only, SSDI only, or both SSI and SSDI						
SSI only	3,389	3,069	93.1	1,730	58.2	54.1
SSDI only	3,139	2,940	94.8	1,748	63.2	60.0
Both SSI and SSDI	1,419	1,323	93.2	814	68.2	63.8
Constructed disability category						
Deaf	85	77	94.1	33	38.8	36.7
Cognitive disability	1,671	1,514	90.9	866	58.9	53.6
Mental illness	2,995	2,762	93.8	1,515	56.6	53.0
Physical disability	3,058	2,861	95.2	1,810	66.2	63.0
Unknown	138	118	88.6	68	68.9	61.3
Beneficiary's age (four categories)						
18 to 29	2,356	2,130	90.6	1,207	57.4	51.9
30 to 39	2,243	2,053	91.6	1,151	57.0	52.1
40 to 49	2,153	2,012	93.5	1,209	60.6	56.7
50 and older	1,195	1,137	95.3	725	64.6	61.6
Sex						
Male	4,206	3,860	93.8	2,187	59.3	55.6
Female	3,741	3,472	94.5	2,105	65.8	62.2
Ethnicity (Hispanic or not)						
Hispanic	346	305	91.7	179	59.2	53.9
Non-Hispanic	7,601	7,027	94.2	4,113	62.6	59.0
Race						
White	3,810	3,527	93.6	2,086	62.5	58.5
Black	1,547	1,421	95.7	823	65.3	62.6
Hispanic	346	305	91.7	179	59.2	53.9
Asian American, Pacific Island American,	72	64	93.8	33	29.7	27.7
American Indian, or Alaska Native	32	32	100.0	19	68.8	68.6
Unknown	2,140	1,983	94.3	1,152	61.0	57.6
Living situation						
Living alone	4,206	3,858	93.3	2,255	61.7	57.6
Living with others	297	265	91.7	167	66.3	61.0
Living with parents	155	132	86.8	54	41.4	36.0
In institution or unknown	68	63	93.2	35	66.0	62.0
Unknown	3,221	3,014	94.8	1,781	63.1	59.9

Table III.3. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Weighted location rate	Count	Weighted cooperation rate	Weighted Response rate
Did the applicant for benefits live in the same ZIP code as the beneficiary?						
No	696	627	93.2	318	49.1	45.8
Yes	3,870	3,553	93.2	2,120	63.7	59.5
No information	3,381	3,152	94.8	1,854	63.2	59.9
Identity of the payee with respect to the beneficiary						
Beneficiary received payments directly	335	309	92.1	167	54.1	50.0
Payee is a family member	2,511	2,279	92.0	1,302	57.3	52.6
Payee is an institution	375	357	94.8	177	54.3	51.6
Other	183	167	91.3	87	62.7	57.2
No information	4,543	4,220	94.7	2,559	64.4	61.0
Number of phone numbers in file						
Only one phone number in file	1,378	1,196	88.6	735	64.8	57.5
Two phone numbers in file	1,946	1,781	93.4	1,046	60.2	56.3
Three phone numbers in file	2,004	1,908	96.7	1,104	60.6	58.6
Four phone numbers in file	1,673	1,571	95.5	889	64.2	61.4
Five or more phone numbers in file	879	822	94.6	472	62.5	59.1
No phones on file, or no information	67	54	94.1	46	85.6	80.5
Number of addresses in file						
One address in file	2,186	2,009	94.0	1,218	61.9	58.2
Two addresses in file	2,264	2,096	93.6	1,202	61.3	57.4
Three addresses in file	1,884	1,748	94.5	1,020	63.0	59.7
Four addresses in file	1,050	955	94.5	552	64.1	60.8
Five or more addresses in file	563	524	93.1	300	63.8	60.1
Census region						
Midwest	1,685	1,569	94.4	991	67.0	63.1
Northeast	1,464	1,357	95.4	740	58.8	56.2
South	3,261	2,999	94.3	1,789	64.3	60.7
West	1,537	1,407	92.1	772	56.6	52.1
Census division						
East North Central	1,153	1,076	94.7	714	69.5	65.7
East South Central	766	697	93.5	437	66.9	62.6
Middle Atlantic	1,046	962	94.9	525	59.7	56.7
Mountain	507	467	92.9	289	66.5	61.8
New England	418	395	96.7	215	56.4	54.8
Pacific	1,030	940	91.7	483	51.8	47.5
South Atlantic	1,540	1,428	94.8	824	62.8	59.6
West North Central	532	493	93.5	277	60.8	56.7
West South Central	955	874	94.2	528	64.6	61.0

Table III.3. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Weighted location rate	Count	Weighted cooperation rate	Weighted Response rate
Metropolitan status of county						
Metropolitan areas with population of 1 million or more	3,615	3,357	94.2	1,883	58.7	55.4
Metropolitan areas with population of 250,000 to 999,999	2,137	1,963	94.8	1,154	61.8	58.7
Metropolitan areas with population of fewer than 250,000	940	860	93.6	530	64.8	60.7
Nonmetropolitan areas adjacent to large metropolitan areas	305	280	93.0	193	74.0	68.6
Nonmetropolitan areas adjacent to medium or small metropolitan areas	654	603	93.2	370	71.3	66.5
Nonmetropolitan areas not adjacent to metropolitan areas	296	269	93.2	162	69.9	65.2
County with low education						
Yes	950	871	92.7	524	62.3	58.0
No	6,997	6,461	94.3	3,768	62.5	59.0
County with recreation-based economy						
Yes	712	647	91.6	350	59.2	54.2
No	7,235	6,685	94.4	3,942	62.8	59.3
Population loss county						
Yes	264	240	95.2	152	63.3	60.2
No	7,683	7,092	94.1	4,140	62.4	58.8
Retirement destination county						
Yes	1,163	1,063	92.8	617	60.6	56.4
No	6,784	6,269	94.4	3,675	62.8	59.3
County with manufacturing-dependent economy						
Yes	669	622	95.5	379	63.6	60.6
No	7,278	6,710	94.0	3,913	62.3	58.6
County with nonspecialized-dependent economy						
Yes	5,339	4,940	94.2	2,907	62.9	59.3
No	2,608	2,392	94.0	1,385	61.5	57.8
County with government-dependent economy						
Yes	864	792	95.6	461	60.5	57.9
No	7,083	6,540	94.0	3,831	62.7	58.9
High poverty county						
Yes	937	858	94.9	523	65.0	61.7
No	7,010	6,474	94.0	3,769	62.1	58.4
High child poverty county						
Yes	1,221	1,120	95.2	660	63.5	60.6
No	6,726	6,212	93.9	3,632	62.3	58.5

Table III.3. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Weighted location rate	Count	Weighted cooperation rate	Weighted Response rate
County racial/ethnic profile[a]						
County with at least 90 percent non-Hispanic white population	692	637	92.9	421	69.5	64.5
County with plurality or majority Hispanic population	657	596	91.1	338	53.9	49.1
County with majority but fewer than 90 percent non-Hispanic white population	3,719	3,429	94.1	1,978	61.7	58.1
County with a racially/ethnically mixed population, no majority group, less than 20 percent American Indian	2,684	2,488	95.3	1,453	63.9	61.0
County with plurality or majority non-Hispanic black population	195	182	92.6	102	58.6	54.2
DCF earnings category[b]						
Beneficiary with monthly DCF earnings above SGA[c] for three consecutive months in 2015 or 2016	376	348	93.2	183	51.6	47.9
Beneficiary with annual DCF earnings above $7,000 in 2015 or 2016	125	114	88.4	69	61.9	54.7
Beneficiary with annual DCF earnings above $2,000 in 2015 or 2016	332	302	90.1	178	67.4	60.8
Beneficiary with annual DCF earnings above $0 in 2015 or 2016	370	344	93.6	192	56.5	53.0
Beneficiary with no annual DCF earnings in 2015 or 2016	6,744	6,224	94.4	3,670	62.8	59.4

Source: NBS Round 6 (the second round of NBS–General Waves).

[a] No beneficiaries were sampled in the sixth county type, that of counties where at least 20 percent of the population was American Indian.

[b] 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 who were in the first or second category, and so on.

[c] Non-blind substantial gainful activity, or $1,090 in 2015, $1,130 in 2016, and $1,170 in 2017.

DCF=Disability Control File

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.2) 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.3 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. It also includes the weighted location rate (using the original sampling weight), the weighted cooperation rate among located sample members (using the location-adjusted sampling weight), and the weighted overall response rate (using the original sampling weight) 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, 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.[57] 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 collapsing was possible. If collapsing was not possible, then we excluded the covariate or interaction from the pool.[58]

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).[59] 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. We relied on the logistic regression procedures in software that

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

[58] 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.

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

accounted for the sample design to make the final selection of covariates (SURVEYLOGISTIC in SAS and RLOGIST in SUDAAN).

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).[60] 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.4, we summarize the variables used in the model as main effects and interactions for locating a sample member. In Table III.5, we summarize the variables used in the model for cooperation among located sample members.

Table III.4. Location logistic propensity model: Representative Beneficiary Sample

Factors in location model
Main effects
AGECAT (AGE CATEGORY)
RACE
REGION (CENSUS REGION)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
DISABILITY (DISABILITY CATEGORY)
CNTYMANUF (MANUFACTURING-DEPENDENT ECONOMY, COUNTY)
CNTYGOV (GOVERNMENT DEPENDENT ECONOMY, COUNTY)
CNTYRET (COUNTY WITH A HIGH PROPORTION OF RETIREES)
CNTYRACE (COUNTY RACIAL/ETHNIC PROFILE)
Two-Factor Interactions
(NONE)

Table III.5. Cooperation logistic propensity model: Representative Beneficiary Sample

Factors in cooperation model
Main effects
AGECAT (AGE CATEGORY)
REGION (CENSUS REGION)

[60] 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.5. *(continued)*

Factors in cooperation model
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
DISABILITY (DISABILITY CATEGORY)
METRO (METROPOLITAN STATUS OF COUNTY)
GENDER
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
LIVING (LIVING SITUATION)
SSI_SSDI (BENEFICIARY IS RECIPIENT OF SSI, SSDI, OR BOTH)
CNTYRACE (COUNTY RACIAL/ETHNIC PROFILE)
CNTYLOWEDUC (LOW EDUCATION COUNTY)
CNTYNONSP (NON-SPECIALIZED DEPENDENT ECONOMY COUNTY)
Two-factor Interactions
(NONE)

The Cox-Snell R-squared is 0.023 (0.063 when rescaled to have a maximum of 1) for the location model and 0.041 (0.056 when rescaled) for the cooperation model.[61] 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, 60.6 percent of pairs are concordant, 37.9 percent of pairs are discordant,[62] and the p-value for the chi-square statistic from the H-L goodness-of-fit test is 0.864.[63] 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 1.01 to 1.54. For the cooperation model, 57.2 percent of pairs are concordant and 42.2 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.479 for the model. The cooperation adjustment from the model, which is calculated as the inverse of the cooperation propensity score, ranged from 1.11 to 4.82. The overall nonresponse adjustment (the product of the location adjustment and the cooperation adjustment) ranged from 1.19 to 5.75.[64]

Among the variables used in the location and cooperation models shown in Tables III.4 and III.5, the number of levels used in the models is often fewer than the number of levels in Table

[61] 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.

[62] 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).

[63] The H-L 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). SUDAAN provides three options for calculating this test; we used the Satterthwaite option. See the SUDAAN User's Manual for details. A hard copy manual is available for Version 9.0 (Research Triangle Institute, 2004), and an online version is available for Version 11.0 (see www.rti.org/sudaan).

[64] 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.

III.3; the levels collapsed for the models are described following the tables. The factors used in the location model included the following:

- **PHONE.** Count of phone numbers in SSA files. There are six levels: (0) no phone numbers on file; (1)-(4) one, two, three, or four phone numbers on file; (5) five or more phone numbers on file.
- **REGION.** Geographic region of beneficiary's place of residence based on U.S. Census regions with three levels: (1) West, (2) South, (3) Midwest and Northeast.
- **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 three levels: (1) mental illness; (2) physical disability (not deafness); (3) deafness, cognitive disability, or disability 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.
- **CNTYGOV.** County with government-dependent economy. There are two levels: (1) a county where 14 percent or more of average annual labor and proprietors' earnings were derived from federal and state government, or 9 percent or more jobs were in federal or state government during 2010–2012, and (2) a county without this attribute.
- **CNTYMANUF.** County with manufacturing-dependent economy: 23 percent or more of the county's average annual labor and proprietors' earnings were derived from manufacturing, or 16 percent or more of jobs were in manufacturing. There are two levels: (1) the county's economy is dependent upon manufacturing, and (2) the county's economy is not dependent upon manufacturing.
- **CNTYRACE.** County racial ethnic profile. There are three levels: (1) county with racially/ethnically mixed population based on 2010 Census, no majority group, (2) county with population that is majority, but less than 90 percent, non-Hispanic white based on 2010 Census, with black and Hispanic percentages less than 20 percent, and (3) other racial/ethnic profile in count**y.**
- **CNTYRET.** Retirement destination county. There are two levels: (1) Number of residents age 60 and older grew by 15 percent or more between 2000 and 2010 censuses due to net migration; and (2) the county does not have this attribute.

Although we attempted to fit interactions in the model, the final selected model did not have any interactions for locating sample members. In Table III.4, we provide the main effects using the variable names listed above. In Appendix D, we provide 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, (2) age 40 to 49, (3) age 18 to 29 or age 50 or older.
- **PHONE.** Count of phone numbers in SSA files. There are four levels: (1) zero or one phone number on file; (2) two phone numbers on file; (3) three phone numbers on file; (4) four or more phone numbers on file.

- **DISABILITY.** Beneficiary's disability category. There are four levels: (1) cognitive disability, (2) deafness, (3) mental illness, (4) physical disability (not deafness) or disability unknown.
- **REGION.** Geographic region of beneficiary's place of residence based on U.S. Census regions with two levels: (1) Midwest, (2) all other regions (South, West, Northeast).
- **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.
- **SSI_SSDI.** Beneficiary title. There are three levels: (1) SSI only, (2) SSDI only, (3) both SSI and SSDI**.**
- **LIVING.** Beneficiary's living situation. There are three levels: (1) beneficiary lives with his or her parents; (2) beneficiary lives with others; (3) beneficiary lives alone, in an institution, or information unknown
- **PDZIPSAME.** Whether the SSI beneficiary and the SSI applicant for benefits lived in the same zip code. There are two levels: (1) beneficiary and applicant lived in different zip codes; (2) beneficiary and applicant lived in same zip code, beneficiary was a recipient of SSDI only, or information unknown.
- **CNTYRACE.** County racial ethnic profile. There are two levels: (1) county with population that is at least 40 percent Hispanic based on 2010 Census, less than 20 percent non-Hispanic black, and less than 50 percent non-Hispanic white; (2) other racial/ethnic profile in count**y**.
- **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) based on average data from the American Community Survey from 2008–2012 and (2) a county without this attribute.
- **CNTYNONSP.** County with nonspecialized-dependent economy. 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.

Once again, although we attempted to fit interactions in the model, the final selected model did not have any interactions for responding sample members. In Table III.5, we provide the main effects using the variable names. In Appendix D, we provide an expanded form of Table III.5, 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 1.07, which was observed with the second 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; beneficiary title, or recipient status (SSI only, SSDI only, and both); and DCF earnings (five categories derived from DCF earnings in 2015 and 2016—the same categories that were used for the RBS nonresponse models). We conducted no trimming after post-stratification.

C. Successful Worker Sample

As noted earlier, we selected the SWS from the Round 6 population of successful workers, a subset of all SSI/SSDI beneficiaries. The sample was selected from seven successive frames, depending upon when the successful worker was identified. In each successive frame, we allocated the sample within two strata defined by beneficiary type (SSDI only, and SSI, which included both SSI only and concurrent beneficiaries). The total number of successful workers identified across the seven frames was 89,936, and the size of each extract ranged from 7,353 (final extract) to 17,594 (third extract).[65] Due to concerns about the number of successful workers in each extract and their distribution across PSUs, we decided to use a dual sample design for all strata. As a result, we supplemented the clustered sample in each extract with a random sample of successful workers from the entire population of successful workers in the same extract.

We selected all respondents in the clustered sample from PSUs, whereas the unclustered sample included successful workers that may or may not have been in the selected PSUs. We therefore organized the unclustered sample into two strata: in the PSU or not in the PSU. In most cases, respondents selected for the in-PSU stratum of the unclustered sample were also in the clustered sample. The weights for such duplicate cases had to be adjusted appropriately to account for a single respondent's appearance in two independent samples. (In the next subsection, we discuss the compositing scheme used to make the needed adjustments.) In addition, if the central office[66] could not resolve the final status of sample members, it treated them differently in the clustered and unclustered samples. For the clustered sample, the central office sent sample cases that they could not resolve by telephone to the field for further follow-up for attempted personal interviews. In the unclustered sample, interviewers made no further attempt to resolve the status of sample members who could not be resolved in the central office. This process is analogous to the accepted practice of subsampling nonrespondents for more intensive effort—in this case, we sent unresolved cases from the clustered sample for field follow-up, but did not follow up unresolved cases in the unclustered sample. When creating

[65] As noted in Section I.B, this total did not include successful workers whose earnings were not yet uploaded to the DCF at the time of extraction due to a lag in the posting of earnings for some beneficiaries. Furthermore, it did include a small number of cases (4,746 out of 89,936) that met the successful work criteria at the time of the initial extraction, but did not meet the criteria for the time period in question in the updated extraction from November 2020. In the later extraction, the actual weighted total number of successful workers was found to be 288,576. We post-stratified the provisional analysis weights to match this total.

[66] The central office is the Mathematica Survey Operations Center.

composite weights (described in the next section), we zeroed out the weights for the cases in the unclustered sample that would have gone to the field had they been in the clustered sample as they were already represented by those in the clustered sample.[67] In Table III.6, we present the final sample sizes for the SWS. This table shows a final released sample of 7,851 cases in the clustered sample and 5,420 in the unclustered sample, for a total of 13,271 sample cases, of which 490 were selected for both the clustered and unclustered samples, and were therefore duplicated across the two samples.[68]

Table III.6. Survey population and initial augmented and final sample sizes, by sampling extracts and strata in the Successful Worker Sample

Data extraction date	Stratum	Population count[a]	Augmented clustered sample	Augmented sample, unclustered	Released clustered sample	Released unclustered sample
12/1/16	SSDI only, in PSUs	1,581	902	129	708	86
12/1/16	SSDI only, not in PSUs	6,058		493		329
12/1/16	All SSI, in PSUs	2,217	1,148	230	871	154
12/1/16	All SSI, not in PSUs	7,203		747		499
1/15/17	SSDI only, in PSUs	1,379	787	128	604	85
1/15/17	SSDI only, not in PSUs	5,306		492		328
1/15/17	All SSI, in PSUs	1,492	804	165	613	110
1/15/17	All SSI, not in PSUs	4,828		533		355
3/1/17	SSDI only, in PSUs	1,725	896	56	689	38
3/1/17	SSDI only, not in PSUs	6,710		219		146
3/1/17	All SSI, in PSUs	2,226	1,027	86	781	57
3/1/17	All SSI, not in PSUs	6,933		268		179
4/15/17	SSDI only, in PSUs	1,388	698	106	532	70
4/15/17	SSDI only, not in PSUs	4,963		378		252
4/15/17	All SSI, in PSUs	1,186	605	107	454	71
4/15/17	All SSI, not in PSUs	3,804		343		228
6/1/17	SSDI only, in PSUs	1,469	743	112	566	75
6/1/17	SSDI only, not in PSUs	5,526		422		281
6/1/17	All SSI, in PSUs	1,594	730	137	557	91
6/1/17	All SSI, not in PSUs	4,886		419		279
7/15/17	SSDI only, in PSUs	1,174	616	128	476	86
7/15/17	SSDI only, not in PSUs	4,566		499		333
7/15/17	All SSI, in PSUs	1,068	465	48	348	32
7/15/17	All SSI, not in PSUs	3,301		147		98
9/1/17	SSDI only, in PSUs	845	499	219	386	146
9/1/17	SSDI only, not in PSUs	3,411		886		591
9/1/17	All SSI, in PSUs	724	350	148	266	98
9/1/17	All SSI, not in PSUs	2,373		484		323
Total	SSDI only, in PSUs	9,562	5,141	878	3,961	586

[67] If a sample member was selected as part of both the clustered and unclustered samples, and the case was sent to the field for further follow-up and was then resolved in the field, the response had to be treated differently between the two samples. For the sample respondent, the value in the clustered sample was recorded according to its final status in the field, whereas the value in the unclustered sample was recorded as "not selected for field follow-up."

[68] The 13,271 released sample cases include 725 that did not meet the criteria for successful work, according to the updated November 2020 extraction.

Data extraction date	Stratum	Population count[a]	Augmented clustered sample	Augmented sample, unclustered	Released clustered sample	Released unclustered sample
Total	SSDI only, not in PSUs	36,540		3,389		2,260
Total	All SSI, in PSUs	10,507	5,129	921	3,890	613
Total	All SSI, not in PSUs	33,330		2,941		1,961
Overall total		**89,936**	**10,270**	**8,129**	**7,851**	**5,420**

[a] The population counts provided here show population totals from the provisional frame from which the sample was drawn. The final population total count was 288,576, as noted earlier in this section.

As indicated, for the clustered samples within each extract, we allocated the sample across the 79 PSUs, with the Los Angeles PSU receiving a double allocation because it had two selections. Given the smaller population sizes for successful workers when compared to the broader beneficiary population, we used only the full PSUs; we did not use the SSUs in the Los Angeles PSU (four SSUs) or the Cook County (Chicago) PSU (two SSUs), which were used for the RBS.

1. Initial weights

We computed the initial weights for the SWS clustered sample based on the probability of selection within the PSU of the augmented sample within the two strata of each extract (SSDI only or SSI) and the probability of selection for the PSU. For the unclustered sample, we computed the initial weights based on the selection probability within the four sampling strata of each extract (SSDI only in PSUs, SSDI only not in any PSU, SSI in PSUs, or SSI not in any PSU). With only a portion of the augmented sample released for use, we then adjusted the initial weights for the sample released for the survey.

2. Dual-frame estimation

To obtain estimates, we had to use a “dual sample design” that combined the clustered and unclustered samples while accounting for different follow-up rules. The design required the creation of composite weights for application to the combined samples. As noted, if the central office could not resolve the final status of a sample member in the unclustered sample, the office determined that the individual was “not selected for field followup” and thus undertook no further efforts to resolve the case. However, if the central office could not resolve the status of a sample member in the clustered sample, the case went to the field for additional data collection efforts (field follow-up).

a. Conceptual framework for composite weights

Consider a survey estimate, *Est(Y)*, such as the proportion of the sample who are currently working, that is computed using information from two independent samples from the same population, such as the clustered and unclustered samples described above. To compute this estimate, the two samples may not be combined without first adjusting the weights because the clustered and unclustered samples in the SWS represent the same target population among successful workers. Separate estimates may be computed from each sample, within each stratum and extract, and then combined by using the following equation:

$$\text{Est}\left(\text{Y}\right)=\lambda \text{Y}_c+\left(1-\lambda\right)\text{Y}_u \quad (1)$$

where Y_c is the survey estimate from the clustered sample for the given payment type, Y_u is the survey estimate from the unclustered sample for the given payment type, and λ is an arbitrary constant between 0 and 1. For example, for successful workers in the first extract in the SSDI only stratum of the Round 6 data, the clustered sample accounted for 275 respondents and the unclustered sample for 122 respondents. The estimates to be combined are the proportion of the 275 in the clustered sample who are currently working and the proportion of the 122 in the unclustered sample who are currently working. In practice, the calculation is more complicated because we need to account for the different rules used in the two samples for following up with nonrespondents or unlocated sample members (discussed later). For the sampling variance, *V(Y)*, the estimate is computed with the following equation:

(2) $$\mathrm{V}(\mathrm{Y}) = \lambda^2 \mathrm{V}(\mathrm{Y}_c) + (1-\lambda)^2\, \mathrm{V}(\mathrm{Y}_u)$$

where $V(Y_c)$ is the sampling variance for the estimate from the clustered sample, and $V(Y_u)$ is the sampling variance for the estimate from the unclustered sample. Any value of λ will result in an unbiased estimate of the survey estimate, but not necessarily an estimate with the minimum sampling variance. To compute the combined-sample estimate with minimum variance, we derive survey estimates by first computing the estimates for each sample, computing a value of λ for each pair of estimates, and then combining the point and variance estimates. While this process produces minimum variance estimates, it is computer-intensive and results in some inconsistencies among estimates for percentages and proportions because of different values of λ among levels of categorical variables. Therefore, since Round 2, we have used an approach that identifies a single lambda calculated by using sample sizes and design effects attributable to unequal weighting for the two samples. In particular, λ acts as a weighting factor, with more weight given to the larger sample. The formula for λ includes sample sizes adjusted for the design effect attributable to unequal weighting. The formula for λ follows:

(3) $$\lambda = \frac{n_c / deff_c}{n_c / deff_c + n_u / deff_u}$$

where n_c and n_u are the sample sizes of the clustered and unclustered central office–located samples, respectively, and $deff_c$ and $deff_u$ are the design effects attributable to unequal weighting for the clustered and unclustered central office–located samples, respectively.

A λ value producing a sampling variance at its minimum value results in the shortest confidence interval and, by implication, the most precise point estimate. A value of lambda that minimizes the variance may be calculated as:

(4) $$\lambda = \mathrm{V}(\mathrm{Y}_u) / \left[\mathrm{V}(\mathrm{Y}_c) + \mathrm{V}(\mathrm{Y}_u)\right]$$

In this case, the minimum variance is:

(5) $$\mathrm{V}(\mathrm{Y}) = \left[\mathrm{V}(\mathrm{Y}_c) * \mathrm{V}(\mathrm{Y}_u)\right] / \left[\mathrm{V}(\mathrm{Y}_c) + \mathrm{V}(\mathrm{Y}_u)\right]$$

b. Application of composite weights to Successful Worker Sample

The population of successful workers may be separated into two parts: the portion requiring field follow-up and the portion not requiring field follow-up. For the latter portion (that is, those whose status was resolved through the central office's data collection efforts), both the clustered and unclustered samples are independent samples that can provide unbiased estimates for this subpopulation. However, for the portion of the target population requiring field follow-up (that is, those whose status was not resolved through the central office's data collection efforts), only the clustered sample can provide unbiased estimates for this subpopulation because unclustered sample cases were not eligible for field follow-up, as it was not selected to be in the clustered sample.

For the subpopulation for which the final status was resolved by the central office, the clustered and unclustered samples may be combined by using the compositing method. The following equation computes the composite weight for each sample member in the clustered central office–resolved sample:

$$WT = \lambda \; WT\left(\text{clustered central office-resolved sample weight}\right) \tag{6}$$

For units in the unclustered central office–resolved sample, the following equation computes the composite weight for each sample member in the unclustered central office–resolved sample:

$$WT = \left(1-\lambda\right) WT\left(\text{unclustered central office-resolved sample weight}\right) \tag{7}$$

Conversely, for the subpopulation of persons whose final status could not be resolved through the central office's data collection efforts, only the clustered sample may be used. In this case, no combining is required, and we used the clustered weight directly as follows:

$$WT = 1 * WT\left(\text{clustered field-resolved sample weight}\right) \tag{8}$$

For unclustered cases that were part of the field-resolved population, the value of the weight is zero. We adjusted the sum of weights among field-resolved cases in the clustered sample so that the total sum matched the original total sum. Given that the weights for each subpopulation (the field-resolved population and the central office-resolved subpopulation) sum to the total number of individuals in each subpopulation, the two subpopulations may simply be combined to form the entire target population.

3. Nonresponse adjustment

As with the Representative Beneficiary Survey, we adjusted the sampling weights in two stages for: (1) sample members who could not be located and (2) sample members who were located and refused to respond. For the SWS, we calculated the nonresponse adjustments (including both the location and cooperation adjustments) by using weighted logistic propensity models, then using the inverse of the propensity score as the weighting adjustment. We treated the extracts (in addition to beneficiary title) as strata in weighting,[69] and calculated the nonresponse adjustments across extracts. We applied the nonresponse adjustments to the composite weights for the clustered and unclustered samples. The result was two weight adjustments, including a location adjustment and a cooperation adjustment, by using logistic propensity models. The models were fitted in the same way as the adjustment models for the RBS (Section B.2 of this chapter).

The main factors or attributes that affected our ability to locate and interview successful worker sample members included the same factors used to locate and interview RBS members: personal characteristics of the sample member (race, ethnicity, gender, and age), identity of the payee with respect to the beneficiary, whether the beneficiary and the applicant for benefits lived in the same location, how many phones or addresses are in the SSA files for the beneficiary, beneficiary's living situation, beneficiary "title" (SSI only, SSDI only, or concurrent), primary disability, and geographic characteristics, including attributes of the county where the beneficiary resides. In subsequent sections, we describe how the specific covariates for each of the weight adjustments varied.

a. Coding of survey dispositions

The scheme used to code respondents included the four general categories described in Section B.2: eligible respondents, ineligible respondents, located nonrespondents, and unlocated sample members.

b. Response rates

The 41.3 percent response rate for the SWS is the product of the weighted location rate and weighted completion rate among located sample members.[70] The weighted location rate is 87.3 percent, and the weighted cooperation rate (the weighted completion rate among located sample members) is 46.9 percent. Analogous to the RBS, we used the weighted rates because the sampling weights vary substantially across the sampling strata, and the weighted rates better reflect the potential for nonresponse bias.

c. Factors related to location and response

In Table III.7, we provide information on selected factors associated with locating a sample member and the factors associated with the response among located sample members. The table includes unweighted counts of all sample members, counts of located sample members, and counts of sample members from whom we obtained a completed interview or whom we deemed ineligible. The table also includes the weighted location rate, weighted cooperation rate among located sample members, and weighted overall response rate for these factors.

[69] In the software that accounted for the sample design, the strata must be identified. The variable that did this was defined according to beneficiary title (SSDI only and SSI) and extract.

[70] Using information from the updated frame from November 2020, the updated weighted SWS response rate was 40.8 percent. This reduction of 0.5 percent was due to the fact that a large percentage of the 725 sampled cases who were not successful workers were found to be ineligible at data collection. Removing these sample cases had a negative effect on the weighted response rate

Table III.7. Weighted location, cooperation, and response rates for Successful Worker Sample, by selected characteristics

	Sample	Located sample		Response among located sample		Overall respondents
	Count[a]	Count	Location rate	Count	Cooperation rate	Response rate[b]
All	**13,271**	**9,842**	**87.3**	**5,050**	**46.9**	**41.3**
Extract						
Extract 1	2,647	1,874	87.1	1,068	48.7	42.6
Extract 2	2,095	1,460	87.4	806	46.7	41.0
Extract 3	1,890	1,535	92.9	842	52.9	49.2
Extract 4	1,607	1,199	92.6	669	49.4	45.8
Extract 5	1,849	1,402	86.3	658	42.5	36.9
Extract 6	1,373	1,045	81.9	474	42.4	34.8
Extract 7	1,810	1,327	75.8	533	38.9	29.7
SSI only, SSDI only, or both SSI and SSDI						
SSI only	3,655	2,680	87.0	1,433	47.5	41.5
SSDI only	6,807	5,091	87.5	2,545	46.7	41.2
Both SSI and SSDI	2,809	2,071	87.5	1,072	46.6	41.2
Constructed disability category						
Deaf	421	290	83.7	117	34.3	28.9
Cognitive disability	1,660	1,160	84.0	582	45.4	38.4
Mental illness	4,913	3,639	87.1	1,811	45.3	39.8
Physical disability	6,142	4,651	88.7	2,478	49.3	44.0
Unknown	135	102	86.0	62	52.4	45.2
Beneficiary's age (four categories)						
18 to 29	3,176	2,240	85.7	1,056	42.2	36.5
30 to 39	3,106	2,281	86.1	1,075	42.6	36.9
40 to 49	2,909	2,143	87.0	1,131	48.4	42.3
50 and older	4,080	3,178	89.9	1,788	53.1	47.9
Sex						
Male	7,131	5,297	87.6	2,580	44.6	39.4
Female	6,140	4,545	87.0	2,470	49.7	43.5
Ethnicity (Hispanic or not)						
Hispanic	610	449	86.4	231	48.8	42.5
Non-Hispanic	12,661	9,393	87.4	4,819	46.8	41.2
Race						
Non-Hispanic White	5,593	4,097	87.5	2,056	46.6	41.1
Non-Hispanic Black	3,535	2,690	87.6	1,417	48.0	42.3
Hispanic	610	449	86.4	231	48.8	42.5
Asian American, Pacific Island American,	127	99	87.7	48	49.7	43.6
American Indian, or Alaska Native	24	17	82.1	9	44.9	38.4
Other or unknown	3,382	2,490	86.9	1,289	45.9	40.2
Living situation						
Living alone	6,016	4,438	87.5	2,346	47.1	41.5
Living with others	385	267	81.7	136	47.6	39.1
Living with parents	37	23	90.7	9	39.6	36.2
In institution or unknown	6,833	5,114	87.5	2,559	46.7	41.2
Did the applicant for benefits live in the same ZIP code as the beneficiary?						
No	816	609	89.4	287	40.1	36.2
Yes	5,540	4,059	86.8	2,177	48.4	42.2
No information	6,915	5,174	87.5	2,586	46.6	41.1

Table III.7. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count[a]	Count	Location rate	Count	Cooperation rate	Response rate[b]
Identity of the payee with respect to the beneficiary						
Beneficiary received payments directly	737	521	84.6	271	48.8	41.8
Payee is a family member	2,325	1,646	84.9	784	42.8	36.6
Payee is an institution	184	138	94.0	60	40.5	38.6
Other	156	112	91.0	58	49.3	45.3
Unknown	9,869	7,425	87.9	3,877	47.8	42.3
Number of phone numbers in file						
Zero or one phone number in file	1,470	964	78.0	545	50.3	39.4
Two phone numbers in file	2,774	1,951	84.5	1,038	49.6	42.3
Three phone numbers in file	3,886	3,006	90.9	1,542	47.0	43.0
Four phone numbers in file	3,586	2,757	88.8	1,359	45.2	40.4
Five or more phone numbers in file	1,555	1,164	88.8	566	43.1	38.7
Number of addresses in file						
Zero or one address in file	2,481	1,819	85.1	986	50.1	43.0
Two addresses in file	3,019	2,244	87.9	1,134	45.9	40.5
Three addresses in file	3,866	2,856	88.1	1,471	47.7	42.4
Four addresses in file	2,643	1,985	87.2	994	44.6	39.2
Five or more addresses in file	1,262	920	88.3	465	45.5	40.4
Census region						
Midwest	2,794	2,022	87.2	1,095	49.7	43.8
Northeast	3,380	2,578	87.8	1,247	44.8	39.5
South	4,025	2,935	86.4	1,562	48.0	41.9
West	3,072	2,307	88.4	1,146	44.6	39.6
Census division						
East North Central	1,979	1,453	87.5	766	49.3	43.7
East South Central	723	546	91.2	289	48.2	44.4
Middle Atlantic	2,269	1,766	88.6	866	45.1	40.1
Mountain	645	465	90.7	254	47.3	42.9
New England	1,111	812	85.8	381	44.1	38.2
Pacific	2,427	1,842	87.5	892	43.5	38.3
South Atlantic	1,973	1,468	87.0	811	50.4	44.2
West North Central	815	569	86.5	329	50.4	43.9
West South Central	1,329	921	82.9	462	44.4	37.1
Metropolitan status of county						
Metropolitan areas with population of 1 million or more	8,242	6,300	87.6	3,133	45.7	40.2
Metropolitan areas with population of 250,000 to 999,999	3,028	2,228	86.9	1,171	47.0	41.1
Metropolitan areas with population of fewer than 250,000	894	587	86.8	333	49.7	43.6
Nonmetropolitan areas adjacent to large metropolitan areas	294	207	91.4	110	41.9	38.7
Nonmetropolitan areas adjacent to medium or small metropolitan areas	485	311	86.5	172	51.3	45.0
Nonmetropolitan areas not adjacent to metropolitan areas	328	209	85.7	131	56.4	49.0

Table III.7. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count[a]	Count	Location rate	Count	Cooperation rate	Response rate[b]
County with low education						
Yes	1,815	1,360	85.5	707	49.4	42.3
No	11,456	8,482	87.6	4,343	46.6	41.1
County with recreation-based economy						
Yes	974	709	87.2	335	40.4	35.4
No	12,297	9,133	87.4	4,715	47.5	41.8
Population loss county						
Yes	634	412	87.1	225	51.3	45.6
No	12,637	9,430	87.4	4,825	46.7	41.0
Retirement destination county						
Yes	1,487	1,072	84.9	528	42.9	36.6
No	11,784	8,770	87.7	4,522	47.4	41.9
County with manufacturing-dependent economy						
Yes	750	509	85.1	270	48.0	41.4
No	12,521	9,333	87.5	4,780	46.8	41.3
County with nonspecialized-dependent economy						
Yes	9,618	7,247	87.6	3,693	47.0	41.5
No	3,653	2,595	86.7	1,357	46.7	40.8
County with government-dependent economy						
Yes	1,542	1,105	88.0	599	48.8	43.2
No	11,729	8,737	87.3	4,451	46.6	41.0
High poverty county						
Yes	1,627	1,188	87.7	625	50.6	61.6
No	11,644	8,654	87.3	4,425	46.4	58.5
County with high level of child poverty						
Yes	1,956	1,455	86.9	766	49.0	41.2
No	11,315	8,387	87.4	4,284	46.6	40.8
Percentage of dwellings that are owner-occupied in county						
Less than 60 percent owner-occupied	4,198	3,129	86.5	1,550	46.9	40.7
Percent owner-occupied between 60 percent and 67.3 percent	4,601	3,507	88.5	1,867	48.4	43.2
Percent owner-occupied exceeds 67.3 percent	4,472	3,206	86.9	1,633	45.6	40.0
County racial/ethnic profile						
County with at least 20 percent American Indian population	22	14	95.8	11	83.7	81.5
County with at least 90 percent non-Hispanic white population	884	592	90.6	337	49.3	45.0
County with plurality or majority Hispanic population	1,387	1,018	86.9	529	48.5	42.3
County with majority but fewer than 90 percent non-Hispanic white population	5,219	3,882	87.0	1,958	45.4	39.8
County with a racially/ethnically mixed population, no majority group, less than 20 percent American Indian	5,290	3,987	86.9	2,027	47.3	41.3
County with plurality or majority non-Hispanic black population	469	349	88.8	188	51.2	44.5

Table III.7. *(continued)*

	Sample	Located sample		Response among located sample		Overall respondents
	Count[a]	Count	Location rate	Count	Cooperation rate	Response rate[b]
DCF earnings category[c]						
Beneficiary with gross annual DCF earnings above $30,000 in 2015 or 2016	2,820	2,069	87.0	949	42.2	37.2
Beneficiary with gross annual DCF earnings above $20,000 in 2015 or 2016	2,855	2,099	87.9	930	44.1	39.0
Beneficiary with gross annual DCF earnings above $15,000 in 2015 or 2016	2,385	1,756	87.6	1,162	51.8	45.7
Beneficiary with gross annual DCF earnings above $7,000 in 2015 or 2016	2,938	2,187	87.9	965	47.4	41.9
Beneficiary with gross annual DCF earnings below $7,000 in 2015 and 2016	2,273	1,731	86.0	1,044	50.4	43.8

Source: NBS Round 6 (the second round of NBS–General Waves).

[a] The sample totals in this column include 725 sample cases that were later found to not meet the criteria for successful work.

[b] Using information from the updated frame from November 2020, the updated weighted SWS overall response rate was 40.8 percent. Other response rates in this table would be similarly reduced.

[c] 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.

d. Propensity models for weight adjustments

The weight adjustments used in the SWS were based on predicted propensities from a logistic regression model. The model-fitting process was similar to that used in the RBS, We identified candidate interactions using CHAID, identified variables to investigate further using the STEPWISE procedure in SAS, then proceeded to create parsimonious models using SURVEYLOGISTIC in SAS, and the RLOGIST procedure in SUDAAN. As indicated earlier, we calculated the adjustments by taking the inverse of the predicted location and cooperation propensities. The adjusted weight for each sample case is the product of the initial sampling weight and the adjustment factor, trimmed to ensure that the impact of outlier weights is minimized.

Tables III.8 and III.9 provide a summary of the variables that were included in the final location and cooperation propensity models. (Appendix D details how the levels were collapsed for each model.)

Table III.8. Location logistic propensity model: Successful Worker Sample

Factors in Location Model
Main Effects
EXTRACT
AGECAT (AGE CATEGORY)
REGION (CENSUS REGION)
BENEFICIARY TITLE (BENEFICIARY OF SSDI, SSI, OR BOTH)
LIVING SITUATION
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
DISABILITY (DISABILITY CATEGORY)
CNTYNONSP (NONSPECIFIC-DEPENDENT ECONOMY, COUNTY)
CNTYGOV (GOVERNMENT DEPENDENT ECONOMY, COUNTY)
CNTYRACE (COUNTY RACIAL/ETHNIC PROFILE)
Two-Factor Interactions
(NONE)

Table III.9. Cooperation logistic propensity model: Successful Worker Sample

Factors in Cooperation Model
Main Effects
EXTRACT
AGECAT (AGE CATEGORY)
REGION (CENSUS REGION)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
DISABILITY (DISABILITY CATEGORY)
EARNINGS CATEGORY
GENDER
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
REPREPAYEE (IDENTITY OF PAYEE WITH RESPECT TO BENEFICIARY)
CNTYREC (COUNTY WITH RECREATION-BASED ECONOMY)
CNTYRACE (COUNTY RACIAL/ETHNIC PROFILE)
Two-Factor Interactions
DISABILITY * AGECAT

The Cox-Snell R-squared is 0.029 (0.055 when rescaled to have a maximum of 1) for the location model and 0.042 (0.056 when rescaled) for the cooperation model.[71] 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, 64.7 percent of pairs are concordant, 34.3 percent of pairs are discordant,[72] and the p-value for the chi-square statistic from the Hosmer-Lemeshow (H-L) goodness-of-fit test is 0.738.[73] 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 1.02 to 1.72. For the cooperation model, 61.4 percent of pairs are concordant and 38.1 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.461 for the model. The cooperation adjustment from the model, which is calculated as the inverse of the cooperation propensity score, ranged from 1.27 to 5.87. The overall nonresponse adjustment (the product of the location adjustment and the cooperation adjustment) ranged from 1.32 to 8.22.

Among the variables used in the location and cooperation models shown in Tables III.8 and III.9, the number of levels used in the models is often fewer than the number of levels in Table III.7; the levels collapsed for the models are described following the tables. The factors used in the location model included the following:

- **EXTRACT.** There are seven levels: (1)-(7) extract number.
- **MOVE.** Count of addresses in SSA files. There are five levels: (1) one address on file; (2)-(4) two, three, or four addresses on file; (5) five or more addresses on file.
- **REGION.** Geographic region of beneficiary’s place of residence based on U.S. Census regions with two levels: (1) West, (2) South, Midwest and Northeast.
- **DISABILITY.** Beneficiary’s disability category. There are two levels: (1) physical disability (not deafness); (2) deafness, mental illness, cognitive disability, or disability 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.

[71] 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.

[72] 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).

[73] 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). SUDAAN provides three options for calculating this test; we used the Satterthwaite option. See the SUDAAN User’s Manual for details. A hard copy manual is available for Version 9.0 (Research Triangle Institute, 2004), and an online version is available for Version 11.0 (see www.rti.org/sudaan).

- **SSI_SSDI.** Beneficiary title. There are two levels: (1) SSDI only, (2) SSI only or both SSI and SSDI.

- **LIVING.** Beneficiary's living situation. There are three levels: (1) beneficiary lives alone; (2) beneficiary lives with others; (3) beneficiary lives with parents, in an institution, or information unknown

- **CNTYGOV.** County with government-dependent economy. There are two levels: (1) a county where 14 percent or more of average annual labor and proprietors' earnings were derived from federal and state government, or 9 percent or more jobs were in federal or state government during 2010–2012, and (2) a county without this attribute.

- **CNTYNONSP.** County with nonspecialized-dependent economy. 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.

- **CNTYRACE.** County racial ethnic profile. There are two levels: (1) county with population that is mostly non-Hispanic white (greater than 90 percent) based on 2010 Census, and (2) other racial/ethnic profile in count**y.**

Although we attempted to fit interactions in the model, the final selected model did not have any interactions for locating sample members. In Table III.8, we provide the main effects using the variable names listed above. In Appendix D, we provide parameter estimates and their standard errors. The factors used in the cooperation model included the following:

- **EXTRACT.** There are seven levels: (1)-(7) extract number.

- **AGECAT.** Beneficiary's age category. There are four levels: (1) age 18 to 29, (2) age 30 to 39, (3) age 40 to 49, or (4) age 50 or older.

- **GENDER.** Beneficiary's sex. There are two levels: (1) male and (2) female.

- **MOVE.** Count of addresses in SSA files. There are five levels: (1) one address on file; (2)-(4) two, three, or four addresses on file; (5) five or more addresses on file.

- **PHONE.** Count of phone numbers in SSA files. There are four levels: (1) zero or one phone number on file; (2)-(4) two to four phone numbers on file; (5) five or more phone numbers on file.

- **DISABILITY.** Beneficiary's disability category. There are four levels: (1) cognitive disability, (2) deafness, (3) mental illness, (4) physical disability (not deafness) or disability unknown.

- **REGION.** Geographic region of beneficiary's place of residence based on U.S. Census regions with three levels: (1) Midwest, (2) South, (3) West or Northeast.

- **REPREPAYEE.** The identity of the payee with respect to the beneficiary. There are two levels: (1) the beneficiary received payments himself or herself; (2) either a family member received benefits on behalf of the beneficiary, an institution received payments on behalf of the beneficiary, or identity of payee not known.

- **PDZIPSAME.** Whether the SSI beneficiary and the SSI applicant for benefits lived in the same zip code. There are two levels: (1) beneficiary and applicant lived in the same zip code; (2) beneficiary and applicant lived in different zip codes, beneficiary was a recipient of SSDI only, or information unknown.

- **EARNCAT.** Earnings category from 2015-2016. There are five levels: (1) gross annual earnings exceeds $30,000 in 2015 or 2016, (2) gross annual earnings never exceeds $30,000 in 2015 and 2016, but exceeds $20,000 in 2015 or 2016, (3) gross annual earnings never exceeds $20,000 in 2015 and 2016, but exceeds $15,000 in 2015 or 2016, (4) gross annual earnings never exceeds $15,000 in 2015 and 2016, but exceeds $7,000 in 2015 or 2016, and (5) gross annual earnings never exceeds $7,000 in 2015 and 2016.

- **CNTYRACE.** County racial ethnic profile. There are three levels: (1) county with racially/ethnically mixed population based on 2010 Census, no majority group, (2) county with population that is majority, but less than 90 percent, non-Hispanic white based on 2010 Census, with black and Hispanic percentages less than 20 percent, and (3) other racial/ethnic profile in county.

- **CNTYREC.** County with recreation-dependent economy. There are two levels: (1) the county's economy is dependent upon recreation, where the indication is determined using three data sources: 1) percentage of wage and salary employment in entertainment and recreation, accommodations, eating and drinking places, and real estate as a percentage of all employment reported by the Bureau of Economic Analysis; 2) percentage of total personal income reported for these same categories by the Bureau of Economic Analysis; and 3) percentage of vacant housing units intended for seasonal or occasional use reported in the 2010 Census; and (2) the county's economy is not dependent upon recreation, or there is no information.[74]

The model also included a single interaction among two of these variables for responding sample members, as noted in Table III.9. In Table III.9, we provide the main effects using the variable names. In Appendix D, we provide an expanded form of Table III.9, with parameter estimates and their standard errors.

4. Trimming

We defined a 14 trimming classes for each model based on beneficiary title (SSDI only and SSI) and the seven extracts. We trimmed 18 weights within these 14 trimming classes. In Table III.10, we present the number of weights trimmed as well as the design effects attributable to unequal weighting before and after trimming for each trimming class, before post-stratification.

[74] The AHRF documentation does not specify the percentage for these three items that will provide an indication that the county has a recreation-dependent economy.

Table III.10. Design effects attributable to unequal weights before and after trimming, within trimming classes in the Successful Worker Sample

			Design effect attributable to unequal weights	
Extract	Sampling stratum	Number of cases trimmed	Before trimming	After trimming
1	SSDI only	3	1.35	1.31
2	SSDI only	1	1.45	1.39 (maximum)
2	SSDI only	4	1.45	1.39
2	SSI	2	1.39	1.31
3	SSDI only	2	1.40	1.39
3	SSI	1	1.40	1.39
4	SSDI only	3	1.49 (maximum)	1.35
4	SSI	0	1.22	1.22
5	SSDI only	0	1.29	1.29
5	SSI	0	1.29	1.29
6	SSDI only	1	1.28	1.28
6	SSI	1	1.26	1.25
7	SSDI only	0	1.22	1.22
7	SSI	0	1.24	1.24

Design effect attributable to unequal weights = $n\Sigma w^2 / \left(\Sigma w\right)^2$

5. Post-stratification

After the nonresponse adjustment and trimming, we post-stratified the weights to the population totals for each extract, and the marginal population totals for three variables obtained from SSA. In particular, the totals were the total number of SSI and SSDI beneficiaries by age (four categories); beneficiary title, or recipient status (SSI only, SSDI only, and both); and DCF earnings (five categories derived from DCF earnings in 2015 and 2016—the same categories that were used for the SWS nonresponse models). We found no extreme weights after post-stratification.

As noted elsewhere in this document (throughout Chapter I, Section III.A.2, and the introduction to Section III.C), the sample was drawn from a provisional frame, which did not match the correct population of successful workers, due to a lag in the posting of earnings for some beneficiaries, or an incorrect provisional posting of earnings for others. Specifically, the provisional frame did not comprise successful workers whose earnings were not included in the DCF at the time of extraction, but did include cases (about 5 percent of the provisional frame) that met the successful work criteria at the time of the initial extraction though should have been excluded, based on an updated extraction from November 2020. In the later extraction, the actual weighted total number of successful workers was found to be 288,576.[75] We post-stratified the

[75] Both of these sample frame counts (89,936 and 288,576) include sampled cases that were found at data collection to be ineligible, either because they had died, were screened out, or were ineligible for other reasons. The later extraction did not check if the beneficiary had become ineligible after the initial extraction date. The weighted estimate of eligible cases using the latest extraction is 265,514.

provisional analysis weights to match this total, matching the marginal totals for age (four categories); beneficiary title, or recipient status (SSI only, SSDI only, and both); DCF earnings (five categories derived from DCF earnings in 2015 and 2016—the same categories that were used for the SWS nonresponse models, but with updated information from November 2020); gender; and disability category (deafness, cognitive disability, mental illness, physical disability, and unknown). We did not match the latest marginal totals for extract.

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.[76]

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.[77] 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

[76] 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.

[77] 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 (.L), 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: (.L) 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, (6) constructed from other variables with imputed values, and (7) longitudinal imputation (using data from an earlier round).[78] 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	8 imputations from SSA's administrative data, 1 longitudinal imputation, 247 imputations from hot deck	256	8,410	3.05
C_Race_i	Race	258 imputations from SSA's administrative data, 1 longitudinal imputation, 469 imputations from hot deck	728	8,410	8.66

Source: NBS Round 6 (the second round of NBS–General Waves).

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. There was one instance where a sample member, a unit respondent in both Rounds 5 and 6, didn't respond to L1 in Round 6, but they did respond to it in Round 5, so we used their Round 5 response. For respondents who still had missing data, we imputed the Hispanic indicator by using a hot deck. The variables used to define the imputation classes for the hot deck depended upon the respondent's surname. We identified those with Hispanic surnames by comparing the respondents' names to those provided by the North American Association of Central Cancer Registries (NAACCR 2003). For those without Hispanic surnames, we defined imputation

[78] Although Round 6 did not include a longitudinal component, there were a small number of individuals who were selected for both the Round 5 and Round 6 samples. A longitudinal imputation is useful if (1) the variable being imputed is one that does not change over time, such as race, and (2) they responded to the question in Round 5 but did not in Round 6.

classes 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 (2016-2017), a file with demographic, health, and economic-related data for every county in the United States. For those with Hispanic surnames, we defined imputation classes by gender and whether the respondent lived in a county where at least 40 percent of the population identified as Hispanic, fewer than 50 percent identified as non-Hispanic white, and fewer than 20 percent identified as non-Hispanic black.

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. There was one instance where a sample member, a unit respondent in both Rounds 5 and 6, didn't respond to L2 in Round 6, but they did respond to it in Round 5, so we used their Round 5 response. 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 44 such cases, 25 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 in three ways: (1) C_MainConBodyGroup_i, which corresponds to the collapsing in Table II.2; (2) C_MainConDiagGrpNewi; and (3) C_MainConColDiagGrp_i. The "New" suffix on C_MainConDiagGrpNew_i is a result of a change in the diagnosis codes that were used in

Round 6. Some of the codes do not map exactly to those used in Round 5.[79] 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_MainConDiagGrpNew_i	Primary diagnosis group	148 hot deck[a]	148	6,968	2.13
C_MainConColDiagGrp_i	Main condition diagnosis group collapsed	148 constructed from imputed variables[a]	148	6,968	2.13
C_MainConBodyGroup_i	Main condition body group	6 hot deck, 142 constructed from imputed variables[a]	148	6,968	2.13
C_DisAge_i	Age at onset of disability	2 longitudinal imputation, 287 hot deck	289	8,410	3.43
C_AdultChildOnset_i	Adult/child onset of disability	28 constructed from imputed variables	28	8,410	0.33
B24_i	Currently working	12 hot deck	12	8,410	0.14

Source: NBS Round 6 (the second round of NBS–General Waves).

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.

[79] For a detailed exposition of the disability codes, see the User's Guide (Callahan, et al. 2019).

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 12 such cases, three of which were imputed as "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 Table IV.3 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, 3 hot deck	4	4,085	0.09
C2_1_1d_i	Main current job SOC code to one digit	9 hot deck[a]	9	4,085	0.22
C8_1_i	Hours per week usually worked at current main job	88 hot deck,[b] 10 imputed by distributional assumptions	98	4,085	2.39
C_TotCurWkHrs_i	Total weekly hours at all current jobs	88 hot deck,[c] 44 constructed from imputed variables	132	4,085	3.23
C_TotCurHrMnth_i	Total hours per month at all current jobs	112 constructed from imputed variables	112	4,085	2.74
C_MainCurJobHrPay_i	Hourly pay at current main job	6 logical, 364 constructed from imputed variables	370	4,084	9.06
C_MainCurJobMnthPay_i	Monthly pay at current main job	62 logical, 22 imputed by distributional assumptions, 342 constructed from imputed variables	426	4,084	10.43
C_TotCurMnthPay_i	Total monthly salary all current jobs	73 logical, 345 hot deck, 32 constructed from imputed variables	450	4,084	11.02

Source: NBS Round 6 (the second round of NBS–General Waves).

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	35 hot deck	35	8,410	0.42
I9_i	Current health	80 hot deck	80	8,410	0.95
I17b_i	Blind or difficulty seeing, even with glasses	3 logical, 96 hot deck	99	8,410	1.18
I19_i	Uses special equipment because of difficulty seeing	1 logical, 19 hot deck, 78 constructed from imputed variables	98	8,410	1.17
I21_i	Deaf or difficulty hearing	1 logical, 94 hot deck	95	8,410	1.13
I22_i	Able to hear normal conversation at all	31 hot deck, 80 constructed from imputed variables	111	8,410	1.32
I23_i	Uses special equipment because of difficulty hearing	15 hot deck, 80 constructed from imputed variables	95	8,410	1.13
I25_i	Difficulty having speech understood	4 logical, 101 hot deck	105	8,410	1.25
I26_i	Able to have speech understood at all	2 logical, 34 hot deck, 78 constructed from imputed variables	114	8,410	1.36
I27_i	Uses special equipment because of difficulty speaking	2 logical, 20 hot deck, 78 constructed from imputed variables	100	8,410	1.19
I29_i	Difficulty walking or climbing stairs without assistance	7 logical, 106 hot deck	113	8,410	1.34
I30_i	Able to walk without assistance at all	67 hot deck, 59 constructed from imputed variables	126	8,410	1.50
I31_i	Uses special equipment because of difficulty walking	44 hot deck, 59 constructed from imputed variables	103	8,410	1.22
I34_i	Able to climb stairs at all	66 hot deck, 59 constructed from imputed variables	125	8,410	1.49
I35_i	Difficulty lifting and carrying 10 pounds	1 logical, 124 hot deck	125	8,410	1.49

Variable name	Description	Imputation method	Number missing	Number eligible	Percentage imputed
I36_i	Able to lift or carry 10 pounds at all	1 logical, 83 hot deck, 92 constructed from imputed variables	176	8,410	2.09
I37_i	Difficulty using hands or fingers	1 logical, 110 hot deck	111	8,410	1.32
I38_i	Able to use hands or fingers at all	38 hot deck, 83 constructed from imputed variables	122	8,410	1.45
I39_i	Difficulty reaching over head	121 hot deck	121	8,410	1.44
I40_i	Able to reach over head at all	1 logical, 47 hot deck, 86 constructed from imputed variables	134	8,410	1.59
I41_i	Difficulty standing	1 logical, 130 hot deck	131	8,410	1.56
I42_i	Able to stand at all	47 hot deck, 75 constructed from imputed variables	122	8,410	1.45
I43_i	Difficulty stooping	1 logical, 115 hot deck	116	8,410	1.38
I44_i	Able to stoop at all	79 hot deck, 64 constructed from imputed variables	143	8,410	1.70
I45_i	Difficulty getting around inside home	3 logical, 107 hot deck	110	8,410	1.31
I46_i	Needs help to get around inside home	22 hot deck, 94 constructed from imputed variables	116	8,410	1.38
I47_i	Difficulty doing errands alone	12 logical, 131 hot deck	143	8,410	1.70
I48_i	Needs help to get around outside home	85 hot deck, 70 constructed from imputed variables	155	8,410	1.84
I49_i	Difficulty getting into/out of bed	2 logical, 114 hot deck	116	8,410	1.38
I50_i	Needs help getting into/out of bed	1 logical, 38 hot deck, 92 constructed from imputed variables	131	8,410	1.56
I51_i	Difficulty bathing or dressing	7 logical, 111 hot deck	118	8,410	1.40
I52_i	Needs help bathing or dressing	33 hot deck, 87 constructed from imputed variables	120	8,410	1.43
I53_i	Difficulty shopping	19 logical, 113 hot deck	132	8,410	1.57

Variable name	Description	Imputation method	Number missing	Number eligible	Percentage imputed
I54_i	Needs help shopping	28 hot deck, 87 constructed from imputed variables	115	8,410	1.37
I55_i	Difficulty preparing own meals	11 logical, 119 hot deck	130	8,410	1.55
I56_i	Needs help to prepare meals	1 logical, 43 hot deck, 87 constructed from imputed variables	131	8,410	1.56
I57_i	Difficulty eating	2 logical, 114 hot deck	116	8,410	1.38
I58_i	Needs help to eat	1 logical, 20 hot deck, 100 constructed from imputed variables	121	8,410	1.44
I59_i	Trouble concentrating or remembering	140 hot deck	140	8,410	1.66
I60_i	Trouble coping with stress	178 hot deck	178	8,410	2.12
I61_i	Trouble getting along with people	160 hot deck	160	8,410	1.90
CageScore_Indicator_i	CAGE Alcohol Score	118 constructed from imputed variables	118	8,410	1.40
I72_i	Uses drugs in larger amounts than prescribed	133 hot deck	133	8,410	1.58

Source: NBS Round 6 (the second round of NBS–General Waves).

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	97 constructed from imputed variables	97	8,410	1.15
C_NumSenLim_i	Number of sensory limitations	131 constructed from imputed variables	131	8,410	1.56
C_NumSevSenLim_i	Number of severe sensory limitations	127 constructed from imputed variables	127	8,410	1.51
C_NumPhyLim_i	Number of physical functional limitations	232 constructed from imputed variables	232	8,410	2.76
C_NumSevPhyLim_i	Number of severe physical functional limitations	272 constructed from imputed variables	272	8,410	3.23
C_NumEmotLim_i	Number of emotional/social limitations	252 constructed from imputed variables	252	8,410	3.00
C_NumADLs_i	Number of impaired ADL	159 constructed from imputed variables	159	8,410	1.89
C_NumADLAssist_i	Number of ADL requiring assistance	159 constructed from imputed variables	159	8,410	1.89
C_NumIADLs_i	Number of IADL difficulties	197 constructed from imputed variables	197	8,410	2.34
C_NumIADLAssist_i	Number of IADL requiring assistance	167 constructed from imputed variables	167	8,410	1.99
C_PCS8TOT_i	Physical summary score	290 constructed from imputed variables	290	8,410	3.45
C_MCS8TOT_i	Mental summary score	290 constructed from imputed variables	290	8,410	3.45
C_DrugDep_i	Drug dependence	137 constructed from imputed variables	137	8,410	1.63

Source: NBS Round 6 (the second round of NBS–General Waves).

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	217 logical, 24 imputed by descriptive statistics using specialized procedures	241	8,410	2.87
C_AmtWorkComp_i	Amount received from workers’ compensation last month	145 logical, 9 imputed by descriptive statistics using specialized procedures	154	8,410	1.83
C_AmtVetBen_i	Amount received from veterans’ benefits last month	133 logical, 16 imputed by descriptive statistics using specialized procedures	149	8,410	1.77
C_AmtPubAssis_i	Amount received from public assistance last month	152 logical, 23 imputed by descriptive statistics using specialized procedures	175	8,410	2.08
C_AmtUnemply_i	Amount received from unemployment benefits last month	135 logical, 5 imputed by descriptive statistics using specialized procedures	140	8,410	1.66
C_AmtPrivPen_i	Amount received from private pension last month	134 logical, 21 imputed by descriptive statistics using specialized procedures	155	8,410	1.84
C_AmtOthReg_i	Amount received from other regular sources last month	136 logical, 18 imputed by descriptive statistics using specialized procedures	154	8,410	1.83

Source: NBS Round 6 (the second round of NBS–General Waves).

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 level of missingness for C_Cohab was considerably higher in Round 6 than in previous rounds, due to a programming error in the software that assigned skip logic in the questionnaire. In particular, all sample members who indicated that they were divorced in question L8 were skipped out of L10, the source variable for C_Cohab. In Rounds 1 through 3 the missingness rate for this variable varied around 0.60%; in Round 4 it increased to 1.02%, and in Round 5 it was 1.26%. This round, it increased to 16.85%, of which 15.19% responded that they were divorced in L8. We were concerned that those who did not respond to C_Cohab because they were divorced would be different than those who did not respond to C_Cohab because they didn't know or refused to respond; therefore, we conducted the imputations among divorced and non-divorced sample members separately. Among divorced cases, 25.8% were imputed to have C_Cohab equal to 1. Among non-divorced cases, 25.4% were imputed to have C_Cohab equal to 1. The reported percentage equal to 1 for this variable was 31.7%.

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[80] 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.

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.

Each of the classing and sorting variables were specific to the variable being imputed.

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

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	417 hot deck	417	8,410	4.96
L3_i	Highest year/grade completed in school	199 hot deck	199	8,410	2.37
L8_i	Marital status	174 hot deck	174	8,410	2.07
L11_i	Living arrangements	10 logical, 159 hot deck	169	8,410	2.01
C_NumChildHH_i	Number of children living in household	8 logical, 149 hot deck, 38 constructed from imputed variables	195	8,410	2.32
C_HHsize_i	Household size	171 hot deck, 33 constructed from imputed variables	204	8,410	2.43
C_Cohab_i	Cohabitation status	565 logical, 851 hot deck	1,416	8,410	16.84
C_FedPovertyLevel_cat	2016 Federal poverty level	3,163 constructed from imputed variables	3,159	8,410	37.56

Source: NBS Round 6 (the second round of NBS–General Waves).

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.[81]

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[82] 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 been developed

[81] 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.

[82] 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).

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.[83] Details about SAS syntax are available from the SAS Institute (2015). Details about SUDAAN syntax are available from RTI International (Research Triangle Institute 2014).

[83] 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.

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 6 Public-Use File Codebook.” Washington, DC: Mathematica Policy Research, 2019.

Callahan, R., K. McDonald, J. Markesich and G. Livermore. “The National Beneficiary Survey-General Waves Round 6 Questionnaire.” Washington, DC: Mathematica, 2019.

Callahan, R., E. Grau, K. McDonald, C. Bush, B. Mory, L. Pranschke, A. Wec and J. Markesich. “The National Beneficiary Survey-General Waves Round 6 (Volume 3 of 3): User’s Guide for Restricted and Public Use Data Files.” Washington, DC: Mathematica Policy Research, 2019.

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 6: Nonresponse Bias Analysis.” Washington, DC: Mathematica Policy Research, 2019.

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.

McDonald, K., B. Mory, R. Callahan, A. Wec, C. Bush, and J. Markesich. “The National Beneficiary Survey—General Waves Round 6: Restricted-Use File Codebook.” Washington, DC: Mathematica, 2019.

McDonald, K., R. Callahan, B. Mory, E. Grau, J. Markesich, A. Wec, and C. Bush. “National Beneficiary Survey—General Waves Round 6 (volume 2 of 3): Data Cleaning and Identification of Data Problems.” Washington, DC: Mathematica, 2019.

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.

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.

APPENDIX A

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

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

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

Question #	Question Text	Current Response Options	Additional Categories Created
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
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

Table A *(continued)*

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 = [You/NAME] cannot get help [you need/he needs/she needs] with tasks you would do at work. This includes having someone help you with things like writing, reading, lifting or reaching.	q=Lack skills r=Cannot find a job/job market is bad

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
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
CP13b1	What was it about [your/NAME's] [main/current] job that might have caused [you/NAME] to have to work less or stop working?	01= Job does not pay enough 02= Job does not offer health insurance benefits 03= Need a different schedule or shift 04= Need time to go to medical appointments 05= Got fired for missing too much time for appointments or hospitalization 06= Health interferes with job performance 07= Do not have the strength, physical energy, or stamina required to work 08= Pain interferes with working a set schedule 09= Personal care and getting ready for work take too long 10= Do not have special equipment or medical devices needed in order to work 11= Other (Specify)	20= Found another job 22= Work schedule 23= Did not like/get along with co-workers 24= Did not like/get along with manager, supervisor, or boss 25= Did not like/get along with other staff responsible for hiring or providing accommodations (such as Human Resources)

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
CP13c1	What was it about [your/NAME's] personal circumstances that might have caused {you/NAME} to have to work less or stop working?	01= Need help caring for children or others 02= Need personal assistance 03= Get injured 04= Might lose benefits such as Social Security, SNAP, Medicaid/Medicare 05= Personality conflicts with others at the job 06= Might get fired for behavior at the job 07= Do not have reliable transportation to and from work 08= Drug/alcohol relapse 09= Would rather do other things than work 10= Do not like working 11= Work is too tiring or stressful 12= Other (Specify)	19= Moved to another area 21= Loss or potential loss of government benefits
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

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
C_BP13b1	What was it about [your/NAME's] [main/current] job that might have caused [you/NAME] to have to work less or stop working?	01= Job does not pay enough 02= Job does not offer health insurance benefits 03= Need a different schedule or shift 04= Need time to go to medical appointments 05= Got fired for missing too much time for appointments or hospitalization 06= Health interferes with job performance 07= Do not have the strength, physical energy, or stamina required to work 08= Pain interferes with working a set schedule 09= Personal care and getting ready for work take too long 10= Do not have special equipment or medical devices needed in order to work 11= Other (Specify)	20= Found another job 22= Work schedule 23= Did not like/get along with co-workers 24= Did not like/get along with manager, supervisor, or boss 25= Did not like/get along with other staff responsible for hiring or providing accommodations (such as Human Resources)
C_BP13c1	What was it about [your/NAME's] personal circumstances that might have caused {you/NAME} to have to work less or stop working?	01= Need help caring for children or others 02= Need personal assistance 03= Get injured 04= Might lose benefits such as Social Security, SNAP, Medicaid/Medicare 05= Personality conflicts with others at the job 06= Might get fired for behavior at the job 07= Do not have reliable transportation to and from work 08= Drug/alcohol relapse 09= Would rather do other things than work 10= Do not like working 11= Work is too tiring or stressful 12= Other (Specify)	19= Moved to another area 21= Loss or potential loss of government benefits
C_B39b	[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?

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
C_B39_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
DP1b_1	What was it about [your/NAME's] job that made [you/him/her] leave it?	01= Job did not pay enough 02= Job did not offer health insurance benefits 03= Needed a different schedule or shift 04= Needed time to go to medical appointments 05= Got fired for missing too much time for appointments or hospitalization 06= Health interfered with job performance 07= Did not have the strength, physical energy, or stamina required to work 08= Pain interfered with working a set schedule 09= Personal care and getting ready for work took too long 10= Did not have special equipment or medical devices needed in order to work 11= Personality conflicted with others at the job 12= Got fired for behavior at the job 13= Other (Specify)	20= Found another job 22= Work schedule 23= Seasonal/Temporary job

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
DP1c_1	What was it about [your/NAME's] personal circumstances that made [you/him/her] leave the job?	01= Need help caring for children or others 02= Need personal assistance to get ready for work each day 03= Get injured 04= Might lose benefits such as Social Security, SNAP, Medicaid/Medicare 05= Do not have reliable transportation to and from work 06= Drug/alcohol relapse 07= Would rather do other things than work 08= Do not like working 09= Increase in income from another source 10= Other (Specify)	19= Moved to another area 21= Loss or potential loss of government benefits
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 2016? (SPECIFY: <OPEN>)	g=Had medical problems/complications
D25_2	What benefits were reduced or ended as a result of [your/NAME's] job in 2016?	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

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created

D26_h	In 2016, 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 2016? (SPECIFY: <OPEN>)
			i=Better health/treatment
			j=More supportive/helpful employer and/or coworker

SS2b_1	What was it about [your/NAME's] job that makes [you/NAME] think [you/he/she] might go back on benefits?	01= Job does not pay enough 02= Job does not offer health insurance benefits 03= Need a different schedule or shift 04= Need time to go to medical appointments 05= Got fired for missing too much time for appointments or hospitalization 06= Health interferes with job performance 07= Do not have the strength, physical energy, or stamina required to work 08= Pain interferes with working a set schedule 09= Personal care and getting ready for work take too long 10= Do not have special equipment or medical devices needed in order to work 11= Other (Specify)	20= Found another job 22= Work schedule 23= Did not like/get along with co-workers 24= Did not like/get along with manager, supervisor, or boss 25= Did not like/get along with other staff responsible for hiring or providing accommodations (such as Human Resources)

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
SS2c_1	What was it about [your/NAME's] personal circumstances that makes [you/NAME] think [you/he/she] might go back on benefits?	01= Need help caring for children or others 02= Need personal assistance 03= Get injured 04= Might lose benefits such as Social Security, SNAP, Medicaid/Medicare 05= Personality conflicts with others at the job 06= Might get fired for behavior at the job 07= Do not have reliable transportation to and from work 08= Drug/alcohol relapse 09= Would rather do other things than work 10= Do not like working 11= Work is too tiring or stressful 12= Other (Specify)	19= Moved to another area 21= Loss or potential loss of government benefits
SB1b_1	What was it about [your/NAME's] job that made [you/NAME] have to go back on benefits?	01= Job does not pay enough 02= Job does not offer health insurance benefits 03= Need a different schedule or shift 04= Need time to go to medical appointments 05= Got fired for missing too much time for appointments or hospitalization 06= Health interferes with job performance 07= Do not have the strength, physical energy, or stamina required to work 08= Pain interferes with working a set schedule 09= Personal care and getting ready for work take too long 10= Do not have special equipment or medical devices needed in order to work 11= Other (Specify)	20= Found another job 22= Work schedule 23= Did not like/get along with co-workers 24= Did not like/get along with manager, supervisor, or boss 25= Did not like/get along with other staff responsible for hiring or providing accommodations (such as Human Resources)

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
SB1c_1	What was it about [your/NAME's] personal circumstances that made [you/NAME] have to go back on benefits?	01= Need help caring for children or others 02= Need personal assistance 03= Get injured 04= Might lose benefits such as Social Security, SNAP, Medicaid/Medicare 05= Personality conflicts with others at the job 06= Might get fired for behavior at the job 07= Do not have reliable transportation to and from work 08= Drug/alcohol relapse 09= Would rather do other things than work 10= Do not like working 11= Work is too tiring or stressful 12= Other (Specify)	19= Moved to another area 21= Loss or potential loss of government benefits
G13	Where did {you/NAME} go to get this training? Please think about all of the places {you/NAME} went in 2016.	01= Vocational rehabilitation agency or {VRSTATE FROM {NAME'S} CURRENT STATE}, 02= Welfare agency or {STATE WELFARE AGENCY NAME/ ACRONYM FROM {NAME'S} CURRENT STATE}, 03= Mental health agency 04= Some other state agency 05= Workforce center or employment/unemployment office, 06= A private business 07= A school or college 08= Some other type of place? (Specify)	9= On the job training (unspecified)
G18	Where did {you/NAME} go to receive these medical services? Please think about all of the places {you/NAME} went in 2016. Did {you/NAME} go to:	01=A clinic or doctor's office 02=A hospital or 03=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

Table A *(continued)*

Question #	Question Text	Current Response Options	Additional Categories Created
G22	Where did {you/NAME} receive this mental health therapy or counseling? Please think about all of the places {you/NAME} went in 2016. Did {you/NAME} go to CIRCLE ALL	01=A mental health agency, 02=A clinic or doctor's office 03=A hospital, 04=Some other type of place? (SPECIFY: <OPEN>)	06=Residential treatment program/facility 07=Rehab center/counseling center/day program 08=Church or religious institution
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

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

APPENDIX B

SOC MAJOR AND MINOR OCCUPATION CLASSIFICATIONS

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

Table 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

Table B *(continued)*

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

Table B *(continued)*

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

Table B *(continued)*

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

Table 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

Table C *(continued)*

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)

Table C *(continued)*

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

Table C *(continued)*

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

Table D.1. Variables in the location logistic propensity model in the Representative Beneficiary Sample

Main effects	Parameter estimate[a]	Standard error
Variables in the location model, Representative Beneficiary Sample		
Count of phone numbers on file (PHONE)		
Zero phones on file	-0.154	0.377
One phone numbers on file	-0.740**	0.296
Two phone numbers on file	-0.166	0.305
Three phone numbers on file	0.507	0.299
Four phone numbers on file	0.164	0.276
Five or more phone numbers on file	Ref. cell	
Beneficiary’s age category (AGECAT)		
Age in range 18 to 29 years	-0.580**	0.181
Age in range 30 to 39 years	-0.550**	0.160
Age in range 40 to 49 years	-0.348*	0.160
Age in range 50 to FRA	Ref. cell	
U.S. Census region (REGION)		
West	-0.292*	0.139
South, Midwest, or Northeast	Ref. cell	
Beneficiary’s race (RACE)		
White	-0.333*	0.151
Not white	Ref. cell	
Beneficiary’s disability category (DISABILITY)		
Mental illness	0.313	0.163
Physical disability (not deafness)	0.441*	0.218
Deafness, cognitive disability, or disability unknown	Ref. cell	
Racial/ethnic profile of county (CNTYRACE)		
County with racially/ethnically mixed population, no majority group	0.411**	0.152
County with majority but less than 90 percent white population	0.456*	0.190
County that doesn’t have this attribute	Ref. cell	
County with manufacturing dependent economy (CNTYMANUF)		
County with manufacturing dependent economy	0.475	0.314
County that doesn’t have this attribute	Ref. cell	
County with government-dependent economy (CNTYGOV)		
County with government-dependent economy	0.447*	0.184
County that doesn’t have this attribute	Ref. cell	
County with high levels of children living in poverty (CNTYRET)		
County with high proportion of retirees	-0.351	0.225
County that doesn’t have this attribute	Ref. cell	
Two-factor interactions[b]		
(none)		

[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.

FRA = full retirement age

Table D.2. Variables in the cooperation logistic propensity model in the Representative Beneficiary Sample

Main Effects	Parameter estimate[a]	Standard error
Variables in the cooperation model, Representative Beneficiary Sample		
Count of phone numbers on file (PHONE)		
Zero or one phone number on file	0.168	0.136
Two phone numbers on file	-0.092	0.147
Three phone numbers on file	-0.108	0.129
Four phone numbers on file	0.039	0.149
Five or more phone numbers on file	Ref. cell	
U.S. Census region (REGION)		
Midwest	0.233*	0.111
Northeast, South, or West	Ref. cell	
Beneficiary’s age category (AGECAT)		
Age in range 30 to 39 years	-0.190*	0.074
Age in range 40 to 49 years	-0.117	0.069
Age in range 18 to 29 years, or 50 to FRA	Ref. cell	
Gender (GENDER)		
Female	0.267**	0.090
Male	Ref. cell	
Beneficiary’s disability category (DISABILITY)		
Mental illness	-0.360**	0.086
Cognitive disability	-0.318**	0.114
Deafness	-1.186**	0.387
Physical disability, excluding deafness, or disability unknown	Ref. cell	
Indicator whether beneficiary and applicant for benefits are in same zip code (PDZIPSAME)		
Applicant and beneficiary live in different zip code	-0.488**	0.133
Applicant and beneficiary live in same zip code, or no information	Ref. cell	
Beneficiary’s living situation (LIVING)		
Beneficiary lives with his or her parents	-0.724**	0.225
Beneficiary lives alone, in an institution, or situation unknown	Ref. cell	
Beneficiary title (SSI_SSDI)		
SSI only recipient	-0.395**	0.124
SSDI only recipient	-0.376**	0.113
Concurrent (recipient of both SSI and SSDI)	Ref. cell	

Table D.2. *(continued)*

Main Effects	Parameter estimate[a]	Standard error
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in nonmetropolitan area not adjacent to metropolitan area	0.574*	0.237
Beneficiary resides in nonmetropolitan area adjacent to medium or small metropolitan area	0.533**	0.103
Beneficiary resides in nonmetropolitan area adjacent to large metropolitan area	0.531**	0.206
Beneficiary resides in metropolitan statistical area (MSA) of less than 250,000	0.318**	0.114
Beneficiary resides in metropolitan statistical area (MSA) of 250,000-999,999	0.175	0.110
Beneficiary resides in metropolitan statistical area (MSA) of 1 million or more	Ref. cell	
Racial/ethnic profile of county (CNTYRACE)		
County that is at least 40% Hispanic, no other majority group	-0.662**	0.103
County that doesn't have this attribute	Ref. cell	
County with non-specialized-dependent economy (CNTYNONSP)		
County with non-specialized-dependent economy	0.235**	0.085
County that doesn't have this attribute	Ref. cell	
County with low levels of education (CNTYLOWEDUC)		
County with low levels of education	0.539**	0.082
County that doesn't have this attribute	Ref. cell	
Two-Factor Interactions[b]		
(none)		

[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

FRA = full retirement age

Table D.3. Variables in the location logistic propensity model in the Successful Worker Sample

Main effects	Parameter estimate[a]	Standard error
Variables in the location model, Successful Worker Sample		
Extract (EXTRACT)		
First extract	Ref. cell	
Second extract	0.013	0.159
Third extract	0.653**	0.156
Fourth extract	0.588**	0.157
Fifth extract	-0.103	0.146
Sixth extract	-0.437**	0.143
Seventh extract	-0.820**	0.117
Count of addresses on file (MOVE)		
One address on file	-0.276	0.150
Two addresses on file	-0.021	0.131
Three addresses on file	-0.035	0.135
Four addresses on file	-0.133	0.138
Five or more addresses on file	Ref. cell	
Beneficiary’s age category (AGECAT)		
Age in range 18 to 29 years	-0.374**	0.109
Age in range 30 to 39 years	-0.388**	0.109
Age in range 40 to 49 years	-0.306**	0.102
Age in range 50 to FRA	Ref. cell	
Beneficiary’s living situation (LIVING)		
Beneficiary lives alone	-0.835	0.534
Beneficiary lives with others	-1.231*	0.566
Beneficiary lives with family, in an institution, or situation unknown	Ref. cell	
U.S. Census region (REGION)		
West	0.192*	0.079
South, Midwest. or Northeast	Ref. cell	
Beneficiary title (SSI_SSDI)		
SSDI only recipient	-0.958	0.529
Recipient of SSI (concurrent or SSI only)	Ref. cell	
Beneficiary’s disability category (DISABILITY)		
Physical disability (not deafness)	0.199*	0.078
Deafness, mental illness, cognitive disability, or disability unknown	Ref. cell	
County with non-specialized dependent economy (CNTYNONSP)		
County with non-specialized dependent economy	0.245*	0.106
County that doesn’t have this attribute	Ref. cell	
County with government-dependent economy (CNTYGOV)		
County with government-dependent economy	0.242	0.125
County that doesn’t have this attribute	Ref. cell	
County with government-dependent economy (CNTYRACE)		
County with population that is 90% white or more	0.514**	0.147
County that doesn’t have this attribute	Ref. cell	
Two-factor interactions[b]		
(none)		

Table D.3. *(continued)*

[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.

FRA = full retirement age

Table D.4. Variables in the cooperation logistic propensity model in the Successful Worker Sample

Main Effects	Parameter estimate[a]	Standard error
Variables in the cooperation model, Successful Worker Sample		
Extract (EXTRACT)		
First extract	Ref. cell	
Second extract	-0.106	0.091
Third extract	0.151	0.083
Fourth extract	-0.010	0.081
Fifth extract	-0.338**	0.088
Sixth extract	-0.352**	0.095
Seventh extract	-0.504**	0.093
Count of phone numbers on file (PHONE)		
Zero or one phone number on file	0.294**	0.107
Two phone numbers on file	0.310**	0.104
Three phone numbers on file	0.166	0.089
Four phone numbers on file	0.076	0.085
Five or more phone numbers on file	Ref. cell	
Count of addresses on file (MOVE)		
One address on file	0.180	0.104
Two addresses on file	0.021	0.099
Three addresses on file	0.050	0.096
Four addresses on file	-0.078	0.096
Five or more addresses on file	Ref. cell	
U.S. Census region (REGION)		
South	0.159	0.099
Midwest	0.214*	0.108
Northeast or West	Ref. cell	
Beneficiary’s age category (AGECAT)		
Age in range 18 to 29 years	-0.873**†	0.097
Age in range 30 to 39 years	-0.471**	0.075
Age in range 40 to 49 years	-0.216**	0.068
Age in range 50 to FRA	Ref. cell	
Gender (GENDER)		
Female	0.164**	0.048
Male	Ref. cell	
Beneficiary’s disability category (DISABILITY)		
Mental illness	-0.464**†	0.097
Cognitive disability	-0.183*	0.092
Deafness	-0.611**	0.152
Physical disability, excluding deafness, or disability unknown	Ref. cell	
Identity of payee relative to beneficiary (REPREPAYEE)		
Beneficiary received payments himself/herself	0.188	0.108
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	0.227**	0.052
Applicant and beneficiary live in different zip code, or no information	Ref. cell	

Table D.4. *(continued)*

Main Effects	Parameter estimate[a]	Standard error
DCF earnings category in 2015-2016 (EARNCAT)		
Beneficiary with gross annual DCF earnings above $30,000 in 2015 or 2016	-0.457**	0.080
Beneficiary with gross annual DCF earnings above $20,000 in 2015 or 2016, but less than or equal to $30,000	-0.368**	0.082
Beneficiary with gross annual DCF earnings above $15,000 in 2015 or 2016, but less than or equal to $20,000	-0.027	0.083
Beneficiary with gross annual DCF earnings above $7,000 in 2015 or 2016, but less than or equal to $15,000	-0.181*	0.078
All other beneficiaries	Ref. cell	
Racial/ethnic profile of county (CNTYRACE)		
County with racially/ethnically mixed population, no majority group	0.131	0.110
County with population that is majority but less than 90% white	0.168	0.107
County that doesn't have these attributes	Ref. cell	
County with recreation-dependent economy (CNTYREC)		
County with recreation-dependent economy	-0.248	0.198
County that doesn't have this attribute	Ref. cell	
Two-Factor Interactions[b]		
AGECAT*DISABILITY		
Age in range 18 to 29 years * Mental illness	0.462**	0.110
Beneficiary missing one or both of these 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

FRA = full retirement age

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

APPENDIX E

SUDAAN AND SAS PARAMETERS FOR NATIONAL ESTIMATES FROM THE NBS-GENERAL WAVES ROUND 6 SAMPLE

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

SUDAAN EXAMPLE

```
PROC DESCRIPT data="SASdatasetname" filetype=sas design=wr;
nest        A_STRATA A_PSU / missunit;
weight      “weight variable” ;
var         “analysis variables” ;
print nsum wsum mean semean deffmean / style=nchs
wsumfmt=f10.0 meanfmt=f8.4 semeanfmt=f8.4 deffmeanfmt=f8.4;
title "NBS National Estimates, SSI and SSDI beneficiaries";
```

SAS EXAMPLE

```
PROC SURVEYMEANS data=”SASdatasetname”;
strata A_STRATA;
cluster A_PSU;
weight “weight variable” ;
var “analysis variables” ;
title “NBS National Estimates, SSI and SSDI successful workers”;
```

WEIGHT VARIABLES USED FOR CROSS-SECTIONAL ESTIMATES

RBS: Wtr6_ben
SWS: Wtr6_sws
Combined samples: Wtr6_com

NEST VARIABLES USED FOR CROSS-SECTIONAL ESTIMATES

A_STRATA

1. Clustered samples for RBS and SWS

a. A_STRATA = 1000 for non-certainty PSUs
b. A_STRATA = 2110 for Los Angeles County certainty PSU, SSDI only, first extract
c. A_STRATA = 2210 for Los Angeles County certainty PSU, SSI, first extract
d. A_STRATA = 3110 for Cook County certainty PSU, SSDI only, first extract
e. A_STRATA = 3210 for Cook County certainty PSU, SSI, first extract

A_STRATA is defined similarly in the clustered sample certainty PSUs for other extracts, where the third digit is replaced by the extract number

2. Unclustered samples for SWS

a. A_STRATA = 4110 for SSDI only, in PSU, first extract
b. A_STRATA = 4210 for SSI, in PSU, first extract
c. A_STRATA = 5110 for SSDI only, not in PSU, first extract

d. A_STRATA = 5210 for SSI, not in PSU, first extract

A_STRATA is defined similarly in the unclustered sample for other extracts, where the third digit is replaced by the extract number

A_PSU

1. Clustered samples for RBS

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

2. Clustered samples for SWS

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

3. Unclustered samples for SWS

   A_PSU=MPRID

NOTES

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

2. Use SUDAAN's SUBPOPN statement to define the subpopulation for which estimates are wanted. In SAS, use the DOMAIN statement

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

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

Mathematica

Princeton, NJ • Ann Arbor, MI • Cambridge, MA
Chicago, IL • Oakland, CA • Seattle, WA
Tucson, AZ • Woodlawn, MD • Washington, DC

EDI Global, a Mathematica Company

Bukoba, Tanzania • High Wycombe, United Kingdom

Mathematica
Progress Together

mathematica.org