National Beneficiary Survey-General Waves Round 7 (Volume 1 of 3): Editing, Coding, Imputation, and Weighting Procedures

Final Report

October 20, 2021

Eric Grau, Yuhong Zheng, Beau Smit, Bevin Mory, Kim McDonald, Ryan Callahan, Hanzhi Zhou, and 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: 40160.324

This page has been left blank for double-sided copying

CONTENTS

Contents

TABLES

Tables

ACRONYMS

ADLs	Activities of daily living
AIC	Akaike's information criterion
CAPI	Computer-assisted personal interviewing
CATI	Computer-assisted telephone interviewing
CHAID	Chi-Squared Automatic Interaction Detector
DCF	Disability Control File
FRA	Full retirement age
IADLs	Instrumental activities of daily living
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 Stator 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_7.html):

- **User's Guide for Restricted- and Public-Use Data Files** (Callahan et al. 2021). 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** (McDonald et al. 2021). 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. 2021). 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 7 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 analysis weights for the representative beneficiary sample; outlines the procedures used to impute missing responses; and discusses procedures that should be used to estimate sampling variances for the NBS.
- **Cleaning and Identification of Data Problems Report** (McDonald et al. 2021). This report describes the data processing procedures performed for Round 7 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. 2021). This report discusses whether the nonresponse adjustments applied to the sampling weights of Round 7 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. 2021). 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 2019, Mathematica conducted the seventh round of data collection since the NBS began in 2004. 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 (2015), 6 (2017), and 7 (2019), we sought to uncover important information about the factors that promote beneficiaries' self-sufficiency and, conversely, the factors that impede beneficiaries' efforts to maintain employment.

For Round 7 of the NBS, we met the goals of the study through three samples: (1) a cross-sectional sample of all beneficiaries (the representative beneficiary sample, or RBS), (2) a cross-sectional sample of a subset of beneficiaries who maintained a minimum level of earnings for a sustained period (a successful worker sample, or SWS), and (3) a subset of SWS cases from Round 6, followed longitudinally in Round 7. The survey was administered to all three of these samples simultaneously. Mathematica collected data by using computer-assisted telephone interviewing (CATI). We deployed in-person field locators to follow-up with some CATI nonrespondents,[1] and we offered computer-assisted personal interviewing (CAPI) to sample members who preferred or needed an in-person interview to accommodate their disabilities.[2]

In this report, we document the editing, coding, weighting, and imputation procedures, as well as the development of standard errors, for Round 7 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, cross-sectional SWS, and longitudinal SWS, and the composite weights that combined weights from the RBS and SWS. In particular, we discuss how we calculated the initial sampling weights, adjusted them to account for nonresponse, used iterative proportional fitting to ensure that weighted marginal totals for selected variables matched frame totals,[3] and trimmed outlier weights when necessary. In

[1] For a portion of the RBS, we did not employ field follow-up. Instead, we randomly selected telephone nonrespondents for a second phase of data collection involving field follow-up, described later in this chapter, in Section B.1. We also did not employ field follow-up for a portion of the SWS. This portion, referred to as the "unclustered" sample, is also described later—in Section B.2 of this chapter.

[2] In Round 7, none of the NBS respondents requested a CAPI interview.

[3] Iterative proportional fitting, or raking, is a method of adjusting weights in an iterative, sequential manner so that weighted marginal totals on key variables of interest match those of the population one variable at a time. It is

Chapter IV, we describe the procedures used to impute missing responses for selected questions and in Chapter V we explain the procedures that should be used to estimate sampling variances for the NBS–General Waves. In Appendix A, we list the open-ended items that were assigned additional categories, as discussed in Chapter II. In Appendices B and C, we list the occupation and industry codes, respectively, which are also discussed in Chapter II. In Appendix D, we provide detailed parameter estimates and standard errors for the weight adjustment models, as discussed in Chapter III. Finally, in Appendix E, we present SUDAAN and SAS parameters that could be used to generate national estimates from the Round 7 sample.[4]

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).

considered a type of post-stratification. For the remainder of this report, we use the terms "raking" and "post-stratification" interchangeably, even though "post-stratification" is a more general term than "raking."

[4] 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)

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 lieu of the Round 5 SWS survey, we conducted in-depth qualitative interviews with 91 successful workers about their experience with benefits and their attempts to find and keep a job (O'Day et al. 2016). In Round 6, we conducted the second cross-sectional survey for the RBS in the NBS–General Waves, using the same primary sampling units (PSUs) that were selected in Round 5, simultaneously conducting the first cross-sectional survey for the SWS. In Round 7, we conducted the third cross-sectional survey for the RBS in the NBS–General Waves,[5] the second cross-sectional survey for the SWS, and a longitudinal follow-up survey for a subset of SWS cases from Round 6. Table I.1 shows the samples that were processed in Rounds 1 through 7.

Table I.1. Summary of Samples Processed in Rounds 1 through 7[a]

Round	Year	Study	RBS	TPS	SWS	Longitudinal 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	√		√	√

[a]Qualitative interviews were also conducted in Round 5 of the NBS-General Waves, in 2015.

Source: NBS Round 7.

B. NBS–General Waves sample design overview

For all survey rounds, the NBS has used a multistage sampling design for both the RBS and cross-sectional SWS, with an independently drawn, supplemental single-stage sample for some successful worker populations.[6] In Round 7, we drew the cross-sectional SWS and RBS independently, from separate frames, although the SWS frame was a subset of the RBS frame. This means that some sample members could have been selected for both the RBS and the cross-sectional SWS—which occurred for 90 individuals (of which 30 responded[7]). Because most

[5] Although this is the third RBS in the NBS–General Waves, it is the seventh RBS over the history of the NBS project.

[6] The RBS and the main sample of the SWS involved selecting individuals within selected clusters of geographic areas, and they are therefore referred to as "clustered samples." 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] Of the 30 who responded, 28 were considered completes for both the cross-sectional SWS and RBS. Of the remaining 2 respondents, 1 was completed in the field for the SWS but was not selected for field operations in the

analyses do not require combining the samples, we did not adjust the RBS and cross-sectional SWS weights for these duplicates. However, in case 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 cross-sectional SWS but could have been potentially sampled for the cross-sectional SWS because they were part of the SWS frame.[8]

The longitudinal SWS was composed of all cases that (1) completed a Round 6 SWS interview and (2) reported currently working at the time of the Round 6 survey.[9] Table I.2 summarizes the actual sample sizes and number of completed interviews for the RBS, cross-sectional SWS, and longitudinal SWS. Note that longitudinal SWS cases carried over from Round 6 also had a chance of being selected, if eligible, for the independently selected Round 7 RBS or the Round 7 cross-sectional SWS.[10]

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 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 Rounds 6 and 7 (for both the RBS and the SWS main sample) that we had selected in Round 5.

1. RBS

For the RBS in Round 7, we fielded a nationally representative sample of 11,299 SSA disability beneficiaries. The sample design for the RBS in Round 7 was similar to the design of the RBS in prior rounds, though there were two important changes: (1) we stratified the sample of PSUs differently in Rounds 1 through 4 than we did in Rounds 5 through 7,[11] and (2) all telephone

second phase of the RBS, and thus was not an RBS complete. The other was an RBS complete but was considered ineligible for the cross-sectional SWS because the person had not been working in the past six months. Therefore, there were 29 total RBS completes, and 29 total cross-sectional SWS completes.

[8] There were an additional 56 sampled cases in the RBS, of which 19 responded, that were part of the SWS frame, but were not sampled for the SWS.

[9] We did not create composite weights that combined sample cases from the longitudinal SWS with any other sample. Longitudinal SWS respondents were selected based on their work activity at Round 6; therefore, they cannot be meaningfully combined with any of the other Round 7 samples.

[10] In general, the only way a longitudinal SWS case would be sent for field follow-up in Round 7 was if it was also selected for one of these other samples and would be sent to the field under those samples' protocols.

[11] As noted earlier, the sample design for Rounds 1 through 4 included two samples: one for all beneficiaries (the RBS) and one for the ticket participants (the TPS). To accommodate the rollout of the TTW program, the PSUs were sampled within strata defined by the three phases of the rollout. The design for Round 5 included one sample only: a

nonrespondents were followed up in the field in Rounds 1 through 6, but only a random sample of telephone nonrespondents were followed up in the field in Round 7, as described in more detail below. 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, 2018.[12] As of that date, the target population consisted of approximately 13.7 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 7 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.

To reduce data collection costs, we implemented a two-phase sample design for the RBS in Round 7. Our goal was to achieve the same number of completed interviews (4,000) as in past rounds, but with a greater proportion completed by phone instead of in the field. In Phase 1, we reserved a minimum of 12 weeks for cases to work their way through the prespecified phone interview protocol for each sample release. Next, in Phase 2, we randomly subsampled telephone nonrespondents for field follow-up instead of fielding all of these cases. This approach necessitated increasing the sample size for the RBS compared with prior rounds. Note that, when weighted for the two-phase design, the weighted response rate is the same regardless of what proportion of Phase 1 nonrespondents is subsampled for Phase 2.

2. Cross-sectional SWS

The cross-sectional 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 nonblind substantial gainful activity (SGA) monthly earnings level ($1,180 in 2018 and $1,220 in 2019) for a minimum of three consecutive calendar months at any time between August 1, 2018, and July 31, 2019, and (2) be younger than age 62 on June 30, 2018.[13] 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

sample of all beneficiaries. The PSUs were not drawn within strata, except those defined by the two certainty PSUs. The Round 6 and Round 7 samples used the same PSUs as those sampled in Round 5.

[12] 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.

[13] We used a 62-year age limit in Round 6 to ensure that longitudinal cases would still be under age 65 at the time of the Round 7 interview. Although we did not plan to follow the Round 7 cross-sectional successful workers longitudinally, we maintained the 62-year age limit in the Round 7 cross-sectional sample for the sake of consistency with Round 6.

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.[14] To ensure a large enough number of 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 7 RBS and are therefore referred to as "extracts" from the larger frame. Using these constraints to define the target population, we identified a population of 101,698 successful workers.[15] Within each of the seven extracts, we stratified the cross-sectional 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 8,590[16] successful workers. We included a screening question as an additional constraint: the sampled successful workers had to indicate that they had been working in the past six months.[17] The targeted number of completed interviews for the two strata was 1,500 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.[18] 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

[14] 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 target population for the sampling effort due to a delay in recording their earnings on the DCF. For these individuals, a lag of up to three years existed between the time that they received their earnings, and the time that the earnings data were recorded in the DCF. There was no way they could be identified in time for the data extraction. Two years after the completion of this document, the DCF earnings data will be revisited, and the weights will be poststratified to account for the new information that the updated DCF earnings data will provide.

[15] This count does not include all beneficiaries who had three consecutive months of earnings above nonblind SGA. It only includes those who met that condition and an additional condition: their earnings were recorded in the DCF at the time of the extraction.

[16] For reasons explained later in this chapter, the cross-sectional SWS includes 152 duplicates in the sample. As a result, 8,438 unique cases were sampled.

[17] 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.

[18] "Historical data" refers to successful worker data from Round 6 as well as earlier simulated extractions of successful workers.

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).[19] We refer to the multistage sample design as the "clustered" sample, and to the second independent sample as the "unclustered" sample.[20] 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 cross-sectional SWS composite weights that combined information from the clustered and unclustered cross-sectional SWS, which appropriately accounted for the different follow-up rules between the two samples.[21] Table I.2 includes the total across the two samples in the cross-sectional SWS, and does not break out the counts between clustered and unclustered samples; the 152 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.

3. Longitudinal SWS

The Round 7 longitudinal sample consists of Round 6 cross-sectional SWS respondents who indicated that they were working at the time of the Round 6 interview. In the Round 6 survey, we defined successful workers as SSI or SSDI beneficiaries who (1) were active or in suspense status due to work[22] on June 30, 2016; (2) had earnings above SSA's nonblind SGA earnings level[23] for at least three consecutive calendar months at any time from August 1, 2016, through July 31, 2017; and (3) were younger than 62 on June 30, 2016. (This is the same definition for successful workers that we used in Round 7, except for the dates and SGA earnings levels.) We used an age limit of 62 to ensure that the longitudinal sample cases would be younger than 65 on the date of the Round 7 interview. Of the 4,587 respondents in the Round 6 SWS, 3,712 were eligible for and included in the Round 7 longitudinal SWS.

[19] 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.

[20] 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).

[21] 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.

[22] "Suspense status due to work" refers to the beneficiaries whose benefits have been temporarily suspended because of work. Those in suspense status for other reasons were not eligible for the sample.

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

Table I.2. NBS–General Waves (RBS and SWS) Round 7 actual sample sizes, target completed interviews, and completed interviews

Sampling strata	Selected sample size	Original target completed interviews[a]	Actual completed interviews
RBS			
Total	11,299	4,000	4,008
18- to 29-year-olds	3,237	1,111	1,127
30- to 39-year-olds	3,291	1,111	1,059
40- to 49-year-olds	3,060	1,111	1,181
50-year-olds or older	1,711	667	704
Cross-sectional SWS			
Total	8,590	3,000	3,017
SSDI only	4,221	1,500	1,493
SSI (SSI only + concurrent)	4,369	1,500	1,524
December 2018 extract	1,757	516	714
SSDI only	833	218	328
SSI (SSI only + concurrent)	924	298	386
January 2019 extract	1,438	456	592
SSDI only	747	222	305
SSI (SSI only + concurrent)	691	234	287
March 2019 extract	1,327	559	446
SSDI only	609	266	207
SSI (SSI only + concurrent)	718	293	239
April 2019 extract	1,043	394	339
SSDI only	545	215	175
SSI (SSI only + concurrent)	498	179	164
June 2019 extract	1,450	444	429
SSDI only	732	230	216
SSI (SSI only + concurrent)	718	214	213
July 2019 extract	998	348	319
SSDI only	468	193	161
SSI (SSI only + concurrent)	530	155	158
September 2019 extract	577	283	178
SSDI only	287	156	101
SSI (SSI only + concurrent)	290	127	77
Longitudinal SWS			
Total	3,712	2,040	2,078
SSDI only	1,863	1,019	1,080
SSI (SSI only + concurrent)	1,849	1,021	998

Source: NBS Round 7.

[a]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, 2015–16, and Round 6 of the NBS. 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. Round 7 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 7 User's Guide (Callahan et al. 2021).

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 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

Source: NBS Round 7.

[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 6 percent only for select

items asking for wages and household income. [24] 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 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 and the case was selected for field follow-up, we deployed a field locator 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 cell phone (or the sample member's own phone, if preferred) to call into the data collection center. If a sample member could not complete the interview by telephone in this manner due to his or her disability, trained field staff were available to conduct the interview in person using CAPI. In Round 7, none of the NBS respondents requested a CAPI interview.

We began the Round 7 CATI data collection in February 2019. In May 2019, Mathematica began in-person locating, which continued concurrently with CATI through November 2019.

1. Completed interviews and response rates

Mathematica completed 9,103 interviews by the end of the Round 7 data collection. Of these, 4,008 were completed from the RBS; 3,017 from the cross-sectional SWS; and 2,078 from the longitudinal SWS. An additional 261 beneficiaries from the RBS, 310 successful workers from the cross-sectional SWS, and 46 longitudinal SWS cases were deemed ineligible for the survey.[25] Because of the independence of the sample selections for the RBS and the cross-sectional SWS, the clustered and unclustered samples within the cross-sectional SWS, and the Round 6 SWS (the source for the Round 7 longitudinal SWS), individuals could be selected for more than one sample. After accounting for 279 cases actually selected for more than one

[24]Item nonresponse was less than 5 percent for the vast majority of variables, but it was 5.01 percent for three constructed disability variables. Details are provided in Chapter IV.

[25] 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 cross-sectional SWS, ineligibles also included sample members who had not worked in the past six months at the time of the interview.

sample, the number of unique completed interviews was 8,824.[26] Mathematica completed all of these interviews by telephone.

The weighted response rates for Round 7 of the NBS are 54.7 percent for the RBS, 41.0 percent for the cross-sectional SWS, and 54.5 percent for the longitudinal SWS.[27] Please see the Round 7 User's Guide (Callahan et al. 2021) for more detailed information on the final survey dispositions.

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 11,299 selected sample cases in the RBS, all 8,590 selected sample cases in the cross-sectional SWS, and all 3,712 cases in the longitudinal SWS to determine if there were systematic differences between respondents and nonrespondents for a variety of covariates. 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. We did find that, after weighting, the estimate of the proportion of the "all others" race category was significantly less than in the frame in the cross-sectional SWS, though this was primarily due to issues other than nonresponse. Any conclusions involving race should be viewed with caution due to the high levels of missing data in that variable.

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 as variables for either sample stratification or post-survey raking, 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 poststratifying them to marginal population totals.

The full nonresponse bias analysis can be obtained from SSA (https://www.ssa.gov/disabilityresearch/nbs_round_7.html).

[26] Among sample cases that were completed interviews only, there were 23 duplicates (46 sample cases total) between the RBS and cross-sectional SWS and 76 duplicates (152 sample cases total) between the clustered and unclustered samples within the cross-sectional SWS. Duplicates and triplicates also occurred with the longitudinal SWS.

[27] Chapter III describes the formulas used to calculate the response rates and alternative formulas that could have been used.

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

II. DATA EDITING AND CODING

Before 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. [2021] 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 computer-assisted telephone interviewing (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 2018, 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."

As part of data processing and based on an initial review of data, we examined verbatim responses to try to find dominant themes for each question. To ensure high quality coding, we used the same coding procedures in Round 7 that we used in prior rounds of data collection. For example, in Round 6, we added supplemental response categories to some field-coded and other/specify response options if a sufficient number of similar responses could not be back-coded into pre-existing categories. In general, we added a new response category to the Round 6 data file if it was provided by 10 percent or more of the respondents who offered a verbatim response to the question. To minimize back-coding during Round 7 data cleaning, we added many of these response categories to the Round 7 instrument. We reused the supplemental response categories that we identified during Round 6 coding, but we did not add to the Round 7 instrument during Round 7 coding.

After reviewing the verbatim responses to the Round 7 open-ended items, we determined that we did not need to add any other response categories to the data file. Appendix A lists all open-ended items that were assigned additional categories during coding.

If the need for changes to the coding scheme became apparent during coding—for example 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. [2021] 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 7.

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 and Round 7. The list of 21 codes used for Round 6 and Round 7 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 biweekly 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 and 7 health 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 nonpsychotic 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

Table II.2 (*continued*)

Code	Label	Description of ICD-9 codes	Corresponding ICD-9 codes
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
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 not elsewhere classified)	800.00–998.9
180	Physical problem, not elsewhere classified)	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 Rounds 6 and 7.

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 2018. 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 Rounds 6 and 7, we used the Bureau of Labor Statistics 2010 Standard Occupational Classification (SOC) for coding.[28] 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,

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

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 and 7 codes applied using the 2010 SOC remain comparable with earlier rounds coded using the 2000 SOC, as all major and minor group 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 Rounds 6 and 7, we used the 2017 NAICS for consistency across rounds.[29] 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. Rounds 6 and 7 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.

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

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.

Source: NBS Round 7.

III. WEIGHTS

We determined the final analysis weights for the RBS, the cross-sectional SWS, and the longitudinal SWS via a three-step process:

1. Calculate the base weights
    a. Calculate initial probability (sampling) weights
    b. Calculate base weights (weights adjusted for two-phase design [RBS] or dual sample design [SWS])
2. Adjust the base weights for two types of nonresponse (location and cooperation)
3. Trim the weights to reduce the variance and the risk associated with outlier weights, and conduct post-survey calibration using raking to ensure weighted marginal totals match frame totals for selected key variables

The initial probability weights are the inverse of the probability of selection and release; the base weights account for peculiarities of the sample design, including the two-phase sampling for the RBS and the dual sampling design for the cross-sectional SWS. In Section A, we summarize the procedures used to compute and adjust the sampling weights. In Sections B, C, and D, respectively, we describe the procedures for computing the weights for the three samples in more detail.

A. Computing and adjusting the weights: A summary

1. RBS

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.[30] 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.[31] These PSUs from Round 5 were used as the first-stage sampling units for Rounds 6 and 7. We selected 79 of these PSUs, with 2 PSUs—Los Angeles County, California, and Cook County, Illinois—acting

[30] 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.

[31] 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.

as certainty PSUs because of their large size.[32] 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.[33] From these SSUs, we selected four SSUs from the Los Angeles County PSU and two from the Cook County PSU.[34] 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, 2018.[35] 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.[36]

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

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

[33] 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 7.

[34] 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.

[35] 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 (3.9 percent) was similar to that of Rounds 5 and 6 (4.0 and 3.9 percent, respectively) but lower than the ineligibility rates obtained in Rounds 1 through 4 (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.

[36] 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).

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 raked to population totals that were obtained from SSA.[37] In this report, these release-adjusted sampling weights are referred to as the base weights. In prior rounds, we released two or three groups of waves, called "releases," after the initial sample (the first release). However, in Round 7, we only released one group of waves after the first release, resulting in only two releases.

As indicated in Chapter I, we used a two-phase sampling procedure for the first time in the Round 7 RBS to increase the proportion of cases completed by phone relative to those completed using field efforts. We used data from Round 6 to project the yield rate among cases sent to the field in the first release. Using this assumed yield rate from Round 6, as well as the phone yield rate in the first release of Round 7, we determined what proportion of second-phase eligible cases (phone nonrespondents) should be randomly selected for the second phase. In the second release, the proportion randomly selected was determined by ensuring that we obtained 4,000 completes. We adjusted the sampling weights of the phone nonrespondents who were selected for the second phase to account for the phone nonrespondents who were not selected to create the final base weights for the RBS.

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 "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

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

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 (2018–2019), 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[38] normalized to the sample size), we made an initial attempt to reduce the pool of covariates and interactions. We used a significance level of 0.3 for entry and retention in the model 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 of variables any covariate or interaction that was clearly unrelated to locating the respondent or to response propensity. We then pooled the variables resulting from the forward and backward procedures as our starting point for the next stage of model fitting.

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.[39] 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.[40] We then used the specific

[38] For the location model, this refers to the probability weight. For the cooperation model, this refers to the location-adjusted probability weight.

[39] 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 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.

[40] 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.

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. 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).

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 need 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. Therefore, the decision to trim requires us to balance increasing bias and decreasing variance. Given our use of the two-phase sample, there was potentially a greater advantage for using trimming to ameliorate the expected increase in the unequal weighting effect. For the RBS, we checked the design effect attributable to unequal weighting within the age-related sampling strata and determined that 64 weights required trimming. The maximum design effect due to weighting among all age strata occurred in the age 30 to 39 stratum, and in the RBS, the effect was reduced by trimming from 1.98 to 1.91.

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,[41] and five categories of annual earnings from the Disability Control Files (DCF) of 2017 and 2018.[42] After post-stratification, we checked the survey weights again to determine

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

[42] 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 post-stratified weights, where the poststratification was based on a variety of earnings categorical variables, each with different cutpoints, some with IRS earnings and some with DCF earnings. We determined that, although the IRS earnings are more accurate than DCF earnings, IRS earnings are only available annually, 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.

2. Cross-sectional SWS

We defined successful workers in the introduction as SSI or SSDI beneficiaries who were (1) active or in suspense on June 30, 2018; (2) with earnings above SSA's nonblind substantial gainful activity (SGA)[43] earnings level for a minimum of three consecutive calendar months at any time between August 1, 2018, and July 31, 2019; and (3) were less than 62 years old on June 30, 2018. 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. 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).[44] 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, 2018; January 15, 2019; March 1, 2019; April 15, 2019; June 1, 2019; and July 15, 2019. 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, 2019 (August 27). 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.

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

have complete information on the amount of earnings that beneficiaries received in that year. For Round 7, we determined five earnings categories using earnings data from the 2017 and 2018 DCF files.

[43] This threshold was $1,170 in 2017 and $1,180 in 2018.

[44] 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.

interview, and how this elapsed time must not exceed six months. The first rectangle corresponds to the first sample extract, which is limited to those whose successful work either ended in October or November in 2018, 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 2018. This is because, for the December extract, we estimated that the successful workers' interview date could be as late as April 2019. 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 2019, 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.

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, 2018	October 2018	May 2019
January 15, 2019	November 2018	June 2019
March 1, 2019	December 2018	July 2019
April 15, 2019	February 2019	September 2019
June 1, 2019	March 2019	October 2019
July 15, 2019	May 2019	December 2019[a]
September 1, 2019	June 2019	January 2020[a]

Source: NBS Round 7.

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

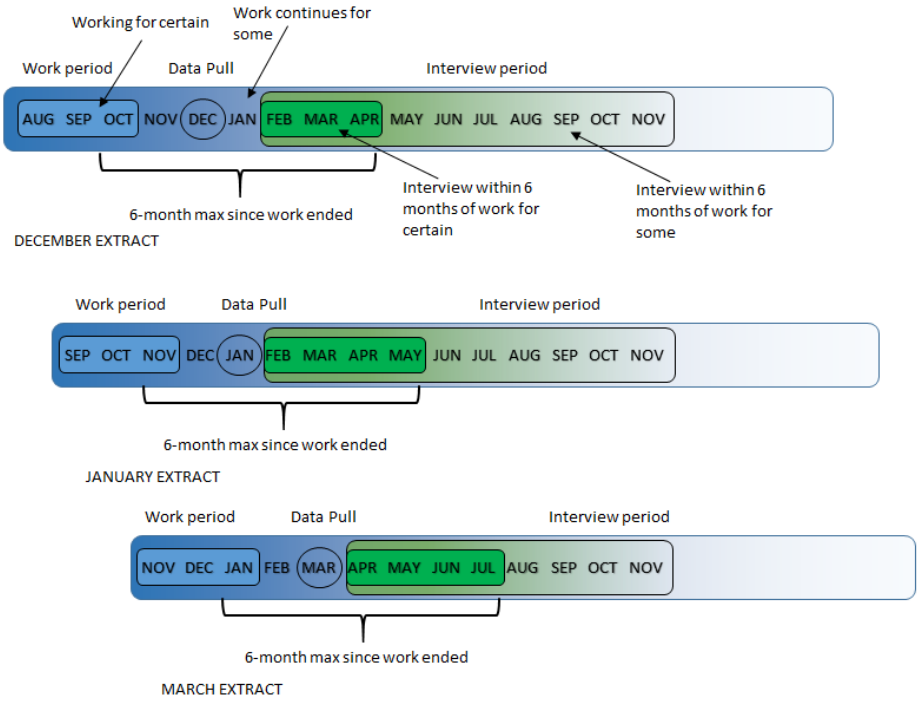
As with the RBS, we used the PSUs as the primary source of sample members for the SWS and selected an initially larger (augmented) sample. We selected the sample of successful workers from among the identified successful workers residing in the same PSUs that were selected for the RBS, and used no SSUs.[45] 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

[45] 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).[46] We refer to the initial sample design as the "clustered" sample; the second independent sample is referred to as the "unclustered" sample.[47] 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 in Section C).

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



Note: The solid rectangles indicate the "for certain" periods, and the gradients represent the decline in certainty over time.

We computed the initial sampling weights for the SWS based on the inverse of the selection probability for the successful worker within each extract, for both the clustered and unclustered samples. 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.)

[46] 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.

[47] 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).

To calculate the base 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 Section C.

We adjusted these base weights for located sample members and then for response among such members. We used logistic propensity models to calculate the location adjustment for all successful workers and the cooperation 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 final adjustment, we adjusted the weights so that the marginal totals matched the frame totals within subgroups defined by five earnings categories,[49] the four age categories, program title,[50] and the extract totals. After post-stratification, we checked the survey again to determine the need for more trimming. Even though the Round 7 weights required trimming before post-stratification in the SWS, they required no further trimming after post-stratification.

3. Longitudinal SWS

As indicated in the introduction, the Round 7 longitudinal SWS consisted of follow-up interviews with a subset of the respondents to the Round 6 cross-sectional SWS. We limited the Round 7 longitudinal sample members to those who, in Round 6, responded affirmatively to question B24 ("Are you currently working at a job or business for pay or profit?"). This restriction removes people who had been working within six months of the Round 6 interview but were not working at the time of the Round 6 interview. The nonresponse-adjusted weights for the Round 6 cross-sectional SWS were used as the "initial probability weights" for the Round 7 longitudinal SWS. As with the Round 7 cross-sectional SWS weights that we summarized in Section A.2, we created Round 7 longitudinal SWS base weights by adjusting the initial probability weights to account for the different follow-up rules for the clustered and unclustered samples in Round 7. This is discussed in Section D of this chapter.

[48] This refers to the creation of weights that combine the unclustered and clustered samples from the SWS. The next section discusses 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.

[49] The five earnings categories used for poststratification 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 2017 and 2018, so the categories were reconfigured to accommodate this.

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

When calculating the nonresponse adjustments, we divided the Round 7 longitudinal sample into two groups, depending on whether the sample members were still SSI or SSDI beneficiaries as of June 30, 2018, and were therefore in the Round 7 beneficiary frame. For both groups, we adjusted for location of the sample members and then for cooperation (response to the survey) among such members. For the group in the Round 7 beneficiary frame—the vast majority of sample members we used logistic propensity models to calculate (1) the location adjustment for all successful workers in the longitudinal sample and (2) the cooperation adjustments for located successful workers in the sample who were current beneficiaries. However, for those who were not in the Round 7 beneficiary frame, we calculated the adjustments using simple weighting classes due to the small number of these members. We created the final weights by trimming and post-stratifying to marginal totals within strata (as the strata were defined when longitudinal SWS cases were originally selected in Round 6), together across the two groups.

4. Composite weights for combining samples (cross-sectional SWS and RBS)

Although the successful worker population constitutes a small subset of the beneficiary population, some studies might require a sample with a substantial number of individuals both within and outside the successful worker population. Such a sample represents a combination of the cross-sectional successful worker and beneficiary samples, requiring the use of another type of composite weight 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. We did not create composite weights that combined sample cases from the longitudinal SWS with any other sample: only the weights from the cross-sectional SWS were used for the composite weights for a combined sample. Sample members in the longitudinal sample were selected based on their work activity at Round 6 and so they cannot be meaningfully combined with any of the Round 7 samples.

In Round 1, some analyses required a combination of data from the RBS and the Ticket Participant Sample, 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.[51] 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 Rounds 6 and 7, 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 45 sample members who happened to be successful workers (whether or not they were actually sampled in

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

the SWS).[52] To ensure representation of the successful worker population, these 45 members of the RBS were represented by the 3,017 members of the SWS who had completed an interview (or had ineligible dispositions after sample selection). The sum of the weights for the 45 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. For the combined weight, we zeroed out the weights for the RBS cases that were also in the SWS frame. We then used a poststratification adjustment so that the weights for the 3,017 responding cases in the SWS added up to the total number of people in the successful worker population, and the weights for the 3,963 non-SWS cases (4,008 − 45) in the RBS added up to the total nonsuccessful worker population.

5. 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 RBS, cross-sectional SWS, longitudinal SWS, and various combinations. For the sake of objectivity, we chose a statistician who was not directly involved in the project.

B. Computing weights for the RBS

1. Base sampling weights

a. Initial probability weights

We computed the initial probability 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,[53] 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 two youngest age strata (18- to 29-year-olds and 30- to 39-year-olds) was 4,500, and for the second-oldest age stratum (40- to 49-year-olds), the sample size was 4,400. The average across these three age groups was roughly four times the target sample size of 1,111, with slightly more cases available in the two youngest age groups, given their historically lower response rates. For beneficiaries age 50 and older, the augmented sample size was 2,600 (just under four times the target sample size of 667). By using the composite size measure already described, we calculated the initial weights for the full augmented sample of 16,000 sample members by taking the inverse of the augmented sampling rate (Fj) for each

[52] Of the 45 successful workers in the RBS, 29 were also part of the SWS.

[53] 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.

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, 2018), initial augmented sample sizes, and initial weights by sampling strata in the NBS

Sampling strata (ages as of June 30, 2018)	Study population	Augmented sample size	Augmented sampling rate (Fj)	Initial sample weights	Released sample
Beneficiaries age 18 to 29	1,346,582	4,500	0.003342	292.44	3,237
Beneficiaries age 30 to 39	1,457,496	4,500	0.003087	323.89	3,291
Beneficiaries age 40 to 49	2,084,746	4,400	0.002111	473.81	3,060
Beneficiaries age 50 to FRA	8,781,834	2,600	0.000296	3377.63	1,711
Total	13,670,658	16,000			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. In Round 7, we released one group of waves after the initial release, for a total of two releases. 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 11,299 beneficiaries (Table III.2).

b. Base weights incorporating two-phase sample design

As described previously, we used a two-phase sample design in the RBS to reduce data collection costs, while maintaining 4,000 completed interviews as we have done in past rounds. We accomplished this by reducing the proportion of completed interviews conducted in the field. Most completed interviews were done in the first phase and were thus conducted by phone, without the need for field follow-up; the second phase involved interviews resulting from field operations.

We defined the first phase of data collection using the typical full set of protocols followed by the central office before we sent a case to the field. According to those protocols, a sample case could be resolved in the first phase if it received a final disposition (such as complete, ineligible, or adamant refusal) without going to the field. Once the protocols for the first phase were exhausted, unresolved cases were eligible for the second phase.

We randomly selected a share of the second-phase eligible cases for further data collection in the field. The decision about how many cases to send to the field was based on a balance between two competing priorities: (1) cost considerations, necessitating fewer cases going to the field,

and (2) precision considerations (achieving the targeted number of completed interviews), necessitating more cases going to the field.

Before collecting data, we assigned a random number between 0 and 1 to each sample case; we used this number in the second phase for any cases that could not be resolved in the first phase. For each of the two sample releases, we set a constant between 0 and 1 and compared it to each second-phase-eligible member's random number to determine whether to send the case to the field. We used data from Round 6 to project the yield rate among cases sent to the field in the first release. Using this yield rate from Round 6, along with the phone yield rate in the first release of Round 7, we determined that 24.4 percent of the phone nonrespondents would be selected for field follow-up in the first release.

If we could not locate and contact a sample member by telephone, we compared their random number to the 0.244 value. For sample members with a random number less than 0.244, we deployed a field locator to make contact in person. Otherwise, we stopped data collection for the case. We used the same procedure for sample members from the second release: the percentage of phone nonrespondents to be randomly selected for field follow-up in this release was 6.0 percent.[54]

Of the 11,299 released cases, 5,030 were resolved in the first phase. For most of these (3,701), the resolved case was a completed interview; however, some cases had other dispositions, such as a final ineligible or adamant refusal, which would have rendered field operations unnecessary. The remaining 6,269 cases were eligible for the second phase, but only 1,128 were selected; of those, only 307 were completed interviews. Therefore, the total number of completed interviews was 3,701 + 307 = 4,008, which is the total observed in Table I.2. We weighted up the 1,128 selected second-phase cases to account for all second-phase eligible cases. For the nonselected second-phase cases, we set the base weights to zero, as they were being represented by the selected cases. Therefore, only 6,158 sample cases (5,030 + 1,128) of the original 11,299 had a positive base weight.[55]

2. Response rates and nonresponse adjustments to the weights

As in virtually all surveys, we had to adjust the base 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,

[54] This small proportion was chosen so that we did not overshoot our desired number of 4,000 completes. However, this created a higher unequal weighting effect than we would have had with a proportion of fielded cases closer to that of the first release.

[55] In Rounds 5 and 6, we selected about 8,000 cases to obtain about 4,000 completes. In Round 7, we needed to select 11,299 cases to obtain 4,000 completes because we would not pursue many of the second-phase-eligible cases in the field, resulting in a lower raw (naïve) yield rate. However, because the second-phase completes have larger base weight, the weighted response rate is the same regardless of the proportion of second-phase eligible cases selected for Phase 2.

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) the number of addresses or phone numbers in the beneficiary's SSA files; (5) the program(s) through which the beneficiary received benefits (SSI, SSDI, or both); and (6) geographic characteristics, including attributes of the county where the beneficiary lived. The following sections detail the steps involved in calculating response rates and adjusting weights for nonresponse.

a. Coding of survey dispositions

The Mathematica 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)[56]

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

4. Unlocated sample members (sample members who could not be located through either central office tracing procedures or in-field searches)

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

b. Response rates

The 54.7 percent response rate for the RBS (Table III.3) is the weighted[57] count of sample members who completed an interview or were deemed ineligible divided by the weighted sample count of all sample members.[58] It can be approximated by taking the product of the weighted location rate and the weighted cooperation rate among located sample members.[59]

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 93 percent (Table III.3). The weighted cooperation rate (that is, the weighted cooperation rate among located sample members) of 58 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.[60] Weighted cooperation rates reflect the rate at which completed interviews are obtained from repeated contact efforts among located persons.

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

[58] 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). Note that the weight used in this calculation is the base weight, already adjusted for the second phase sample selection. 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.

[59] 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.

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

Table III.3. Weighted location, cooperation, and response rates for the RBS, 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	6,158	6,004	93.4	4,269	58.4	54.7
SSI only, SSDI only, or both SSI and SSDI						
SSI only	2,492	2,417	91.6	1,700	55.0	50.4
SSDI only	2,556	2,500	93.9	1,782	60.0	56.5
Both SSI and SSDI	1,110	1,087	95.5	787	59.2	56.5
Constructed disability category						
Deaf	35	34	95.2	27	67.7	64.0
Cognitive disability	1,187	1,152	91.6	819	59.7	54.9
Mental illness	2,268	2,208	93.0	1,523	56.2	52.4
Physical disability	2,552	2,500	93.9	1,822	59.0	55.5
Unknown	116	110	93.7	78	63.9	60.7
Beneficiary's age						
18 to 29	1,695	1,652	92.8	1,191	56.0	51.9
30 to 39	1,661	1,613	91.8	1,129	55.5	51.0
40 to 49	1,709	1,664	91.9	1,188	55.2	50.7
50 and older	1,093	1,075	94.1	761	60.0	56.6
Sex						
Male	3,225	3,149	92.6	2,130	56.5	52.4
Female	2,933	2,855	94.2	2,139	60.4	57.0
Ethnicity						
Hispanic	222	213	93.4	155	68.1	63.8
Non-Hispanic	5,936	5,791	93.4	4,114	58.1	54.4
Race						
White	3,133	3,061	92.2	2,184	59.1	54.7
Black	1,148	1,116	94.4	805	58.5	55.3
Hispanic	222	213	93.4	155	68.1	63.8
Asian American, Pacific Island American,	60	60	100.0	33	45.1	45.0
American Indian, or Alaska Native	16	13	66.9	6	33.9	23.5
Unknown	1,579	1,541	96.1	1,086	55.2	53.0
Living situation						
Living alone	3,130	3,045	92.7	2,146	56.0	51.9
Living with others	268	263	95.3	203	64.7	61.5
Living with parents	112	108	91.6	76	49.3	45.3
In institution or unknown	52	52	100.0	35	70.8	71.1
Unknown	2,596	2,536	93.8	1,809	59.8	56.3
Did the applicant for benefits live in the same zip code as the beneficiary?						
No	483	467	92.2	312	55.4	51.6
Yes	2,943	2,868	92.9	2,062	56.7	52.6
No information	2,732	2,669	93.9	1,895	59.8	56.3

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
Identity of the payee with respect to the beneficiary						
Beneficiary received payments directly	246	237	95.2	174	62.6	59.4
Payee is a family member	2,041	2,003	94.6	1,423	58.5	55.5
Payee is an institution	253	247	91.4	154	56.9	52.4
Other	116	113	97.6	74	44.8	43.6
No information	3,502	3,404	93.1	2,444	58.7	54.7
Number of phone numbers in file						
One	1,399	1,363	92.8	970	57.5	53.5
Two	1,855	1,810	91.5	1,282	60.2	55.2
Three	1,471	1,437	96.2	997	56.4	54.3
Four	936	916	95.5	670	63.7	60.9
Five or more	415	402	90.1	299	53.2	47.9
Zero, or no information	82	76	74.8	51	34.4	25.0
Number of addresses in file						
One	1,663	1,633	93.4	1,193	62.3	58.3
Two	1,622	1,581	94.7	1,106	59.4	56.2
Three	1,482	1,439	92.4	1,007	55.8	51.6
Four	849	829	92.9	597	61.3	57.2
Five or more	480	463	94.0	329	53.0	50.2
Zero, or no information	62	59	93.9	37	23.3	21.7
Census region						
Midwest	1,337	1,309	93.7	953	59.7	55.9
Northeast	1,121	1,095	91.2	757	55.9	51.3
South	2,516	2,447	94.2	1,779	61.3	57.9
West	1,184	1,153	93.5	780	52.3	48.9
Census division						
East North Central	926	908	93.6	670	61.4	57.3
East South Central	573	562	96.1	413	63.6	61.3
Middle Atlantic	813	790	89.3	539	55.8	50.1
Mountain	407	398	93.2	284	58.9	54.9
New England	308	305	96.3	218	56.4	54.3
Pacific	777	755	93.6	496	48.9	45.8
South Atlantic	1,196	1,163	93.4	826	59.2	55.4
West North Central	411	401	93.9	283	55.6	52.5
West South Central	747	722	94.0	540	63.0	59.1
Metropolitan status of county						
Metropolitan areas with population of 1 million or more	2,778	2,702	93.1	1,854	55.8	52.1
Metropolitan areas with population of 250,000 to 999,999	1,676	1,637	94.9	1,188	55.4	52.7
Metropolitan areas with population of fewer than 250,000	741	727	93.7	529	61.4	57.7

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
Nonmetropolitan areas adjacent to large metropolitan areas	224	218	87.9	175	79.5	69.9
Nonmetropolitan areas adjacent to medium or small metropolitan areas	529	520	94.2	372	65.8	62.2
Nonmetropolitan areas not adjacent to metropolitan areas	210	200	87.4	151	63.6	55.3
County with low education level						
Yes	757	742	97.3	512	58.5	57.0
No	5,401	5,262	92.9	3,757	58.4	54.4
County with recreation-based economy						
Yes	558	538	91.0	370	64.0	58.7
No	5,600	5,466	93.6	3,899	57.9	54.3
County with population loss						
Yes	220	212	91.2	154	68.9	62.9
No	5,938	5,792	93.5	4,115	58.0	54.4
Retirement destination county						
Yes	902	877	96.5	611	59.9	57.8
No	5,256	5,127	92.9	3,658	62.8	54.1
County with manufacturing-dependent economy						
Yes	537	525	88.8	374	64.8	57.9
No	5,621	5,479	93.9	3,895	57.8	54.4
County with nonspecialized-dependent economy						
Yes	4,156	4,058	94.3	2,887	57.2	53.9
No	2,002	1,946	91.6	1,382	61.1	56.3
County with government-dependent economy						
Yes	642	626	92.9	449	56.7	52.9
No	5,516	5,378	93.5	3,820	58.6	54.9
High poverty county						
Yes	711	691	93.9	508	60.5	56.9
No	5,447	5,313	93.3	3,761	58.1	54.4
High child poverty county						
Yes	931	899	94.7	663	66.4	63.1
No	5,227	5,105	93.2	3,606	57.0	53.2
County racial/ethnic profile[a]						
At least 90 percent non-Hispanic White	530	519	90.4	386	61.6	55.5
Plurality or majority Hispanic	519	500	94.5	339	53.4	50.4
Majority but less than 90 percent non-Hispanic White	2,915	2,844	92.7	1,997	57.5	53.5
Racially/ethnically mixed, no majority group	1,981	1,938	95.1	1,397	59.3	56.4
Plurality or majority non-Hispanic Black	213	203	91.0	150	64.7	59.6

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
DCF earnings category[b]						
Monthly DCF earnings above SGA[c] for three consecutive months in 2017 or 2018	313	305	92.0	196	43.7	41.2
Gross annual DCF earnings above three times SGA in 2017 or 2018	281	274	91.8	200	64.1	58.7
Gross annual DCF earnings above $0 in 2017 or 2018	408	394	93.6	299	64.3	60.3
No annual DCF earnings in 2017 or 2018	5,156	5,031	93.5	3,574	58.6	54.9

Source: NBS Round 7.

[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]Nonblind substantial gainful activity, or $1,170 in 2017, $1,180 in 2018, and $1,220 in 2019.

DCF = Disability Control File.

The sample count in Table III.3 excludes second-phase-eligible cases that were not selected for the second phase, as these cases have zero weight. We used the weighted rates because (1) with two-phase sampling, the unweighted rates are not meaningful;[61] (2) the sampling rates—and thus the sampling weights—vary substantially across the sampling strata (as seen in Table III.2); and (3) 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 cooperation

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 base weight), the weighted cooperation rate among located sample members (using the location-adjusted base weight), and the weighted overall response rate (using the original base weight) for these factors, which helped inform the decision about the final set of variables to be used in the nonresponse adjustment models.

[61] If we included the second-phase-eligible cases that were not selected for the second phase, the unweighted response rate would be too low, and it would not reflect the fact that the cases' base weights were transferred to other sample members. If we excluded these cases, the unweighted response rate would be too high, and it would not reflect the unsuccessful effort to get a response from these cases in the first phase.

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.[62] 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.[63]

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).[64] 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 accounted for the

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

[63] 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.

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

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).[65] 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: RBS

Factors in location model
Main effects
AGECAT (AGE CATEGORY)
RACE
SSI_SSDI (BENEFICIARY TITLE: RECIPIENT OF SSI AND/OR SSDI)
DIVISION (CENSUS DIVISION)
REPREPAYEE (IDENTITY OF PAYEE WITH RESPECT TO BENEFICIARY)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
CNTYRET (COUNTY WITH A HIGH PROPORTION OF RETIREES)
Two-Factor Interactions
(NONE)

Source: NBS Round 7.

[65] 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. Cooperation logistic propensity model: RBS

Factors in cooperation model
Main effects
AGECAT (AGE CATEGORY)
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
REPREPAYEE (IDENTITY OF PAYEE WITH RESPECT TO BENEFICIARY)
GENDER
ETHNICITY (HISPANIC OR NOT)
EARNINGS CATEGORY
METRO (METROPOLITAN STATUS OF COUNTY)
CNTYPERSPOV (COUNTY WITH PERSISTENT HIGH LEVELS OF POVERTY)
CNTYCHPOV (COUNTY WITH PERSISTENT CHILD POVERTY)
CNTYREC (COUNTY WITH RECREATION-BASED ECONOMY)
Two-factor Interactions
CNTYPERSPOV * AGECAT

Source: NBS Round 7.

The Cox-Snell R-squared is 0.028 (0.074 when rescaled to have a maximum of 1) for the location model and 0.035 (0.048 when rescaled) for the cooperation model.[66] 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, 53.5 percent of pairs are concordant, 43.7 percent of pairs are discordant,[67] and the p-value for the chi-square statistic from the H-L goodness-of-fit test is 0.894.[68] Although the percentages that are concordant and discordant are slightly less favorable than in prior rounds, the other diagnostic values indicate a reasonably good fit of the model to the data. The location adjustments from the model, calculated

[66] 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.

[67] 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).

[68] 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).

as the inverse of the location propensity scores, ranged from 1.00 to 1.79. For the cooperation model, 54.1 percent of pairs are concordant and 44.5 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.744 for the model. The cooperation adjustments from the model, which are calculated as the inverse of the cooperation propensity score, ranged from 1.14 to 4.78. The overall nonresponse adjustments (the product of the location adjustment and the cooperation adjustment) ranged from 1.16 to 5.57.[69]

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 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 five levels: Levels 1 through 4 indicate one, two, three, or four phone numbers on file, respectively, and Level 5 indicates no phone numbers or five or more phone numbers on file.
- **DIVISION.** Geographic region of beneficiary's place of residence based on U.S. Census divisions, with two levels: (1) Middle Atlantic division and (2) all other census divisions in the United States.
- **RACE.** Race of beneficiary. There are three levels: (1) non-Hispanic White; (2) non-Hispanic Black; and (3) neither non-Hispanic White nor non-Hispanic Black, or race not known.
- **REPREPAYEE.** The identity of the payee with respect to the beneficiary. There are two levels: (1) a family member received benefits on behalf of the beneficiary, and (2) the beneficiary received payments himself or herself, an institution received payments on behalf of the beneficiary, or the payee's identity is not known.
- **AGECAT.** Beneficiary's age category. There are three levels: (1) age 18 to 29, (2) age 30 to 39, and (3) age 40 or older.
- **GENDER.** Beneficiary's sex. There are two levels: (1) male and (2) female.
- **SSI_SSDI.** Beneficiary title. There are two levels: (1) recipient of SSI only and (2) recipient of SSDI, either with SSI (concurrent) or SSDI only.
- **CNTYRET.** Retirement destination county. There are two levels: (1) the number of residents age 60 and older grew by 15 percent or more between the 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. Table III.4 shows the main effects using the variable

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

names listed above. Appendix D provides the parameter estimates and their standard errors. The factors used in the cooperation model included the following:

- **AGECAT.** Beneficiary's age category. There are four levels: (1) age 18 to 29, (2) age 30 to 39, (3) age 40 to 49, and (4) age 50 or older.
- **MOVE.** Count of addresses in SSA files. There are five levels: Levels 1 through 4 indicate one, two, three, or four addresses on file, respectively, and Level 5 indicates no addresses or five or more addresses on file.
- **ETHNICITY.** Ethnicity of beneficiary. There are two levels: (1) Hispanic and (2) not Hispanic.
- **METRO.** Metropolitan status of beneficiary's county of residence. There are three levels: (1) beneficiary lived in metropolitan area with population of 250,000 or more; (2) beneficiary lived in metropolitan area with population of fewer than 250,000; and (3) beneficiary lived in nonmetropolitan area.
- **GENDER.** Beneficiary's sex. There are two levels: (1) male and (2) female.
- **EARNCAT.** Earnings category from 2017 to 2018. There are four mutually exclusive levels: (1) gross annual earnings exceed SGA for three consecutive months at least once in 2017 or 2018; (2) not in Group 1, but gross annual earnings exceed three times SGA in 2017 or 2018; (3) not in Groups 1 or 2, but gross annual earnings exceed zero in 2017 or 2018; and (4) gross annual earnings are zero in both 2017 and 2018.
- **CNTYREC.** County with recreation-dependent economy. There are two levels. Level 1 indicates that the county's economy depends on recreation, with the indication 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 as reported in the 2010 census. Level 2 indicates that either the county's economy does not depend on recreation or there is no information.[70]
- **CNTYPERSPOV.** County with persistent high levels of poverty. There are two levels. Level 1 indicates a county where 20 percent or more of residents were poor, as measured by the 1980, 1990, and 2000 censuses and the American Community Survey's five-year average data for 2007–11. Level 2 indicates a county without this attribute.
- **CNTYCHPOV.** County with persistent high levels of child poverty. There are two levels. Level 1 indicates that 20 percent or more of county-related children under 18 were poor, as

[70] The Area Health Resource File documentation does not specify the percentage for these three items that would indicate that the county has a recreation-dependent economy.

measured in the 1980, 1990, and 2000 censuses and the American Community Survey's five-year average data for 2007–11. Level 2 indicates a county without this attribute.

The model also included a single interaction, that of CNTYCHPOV by AGECAT. 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.

3. Poststratification and trimming

After we applied adjustments to the base weights, we reviewed the distribution of weights to determine the need for further weight trimming. With the two-phase design, we expected that trimming (within age group) would be needed to ameliorate the increased unequal weighting effect. We trimmed 64 weights to reduce the maximum design effect attributable to unequal weighting from 1.98 to 1.91, which we observed with the second-youngest age stratum.

Poststratification 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 poststratified 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 2017 and 2018—the same categories that were used for the RBS nonresponse models). We conducted no trimming after poststratification.

C. Cross-sectional SWS

As noted earlier, we selected the cross-sectional SWS from the Round 7 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 101,698, and the size of each extract ranged from 8,572 (final extract) to 19,852 (first extract). 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[71] 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 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.[72] In Table III.6, we present the final sample sizes for the SWS. This table shows a final released sample of 6,022 cases in the clustered sample and 2,568 in the unclustered sample, for a total of 8,590 sample cases, of which 152 were selected for both the clustered and unclustered samples, and were therefore duplicated across the two samples.

Table III.6. Survey population and initial augmented and final sample sizes, by sampling extracts and strata in the cross-sectional SWS

Data extraction date	Stratum	Population count	Augmented clustered sample	Augmented sample, unclustered	Released clustered sample	Released unclustered sample
12/1/18	SSDI only, in PSUs	1,815	773	72	588	48
12/1/18	SSDI only, not in PSUs	7,363		295		197
12/1/18	All SSI, in PSUs	2,498	927	80	697	53
12/1/18	All SSI, not in PSUs	8,176		261		174
1/15/19	SSDI only, in PSUs	1,688	641	83	488	55
1/15/19	SSDI only, not in PSUs	6,259		306		204
1/15/19	All SSI, in PSUs	2,019	805	31	607	21
1/15/19	All SSI, not in PSUs	6,221		94		63
3/1/19	SSDI only, in PSUs	1,581	664	28	517	18
3/1/19	SSDI only, not in PSUs	6,300		109		74
3/1/19	All SSI, in PSUs	2,074	774	49	582	33
3/1/19	All SSI, not in PSUs	6,510		155		103
4/15/19	SSDI only, in PSUs	1,434	543	40	411	27
4/15/19	SSDI only, not in PSUs	5,736		160		107
4/15/19	All SSI, in PSUs	1,157	212	120	147	80
4/15/19	All SSI, not in PSUs	3,908		407		271
6/1/19	SSDI only, in PSUs	2,008	752	51	562	35
6/1/19	SSDI only, not in PSUs	7,849		202		135

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

[72] 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."

Table III.6 (*continued*)

Data extraction date	Stratum	Population count	Augmented clustered sample	Augmented sample, unclustered	Released clustered sample	Released unclustered sample
6/1/19	All SSI, in PSUs	1,738	644	83	482	55
6/1/19	All SSI, not in PSUs	5,695		272		181
7/15/19	SSDI only, in PSUs	1,261	476	34	356	22
7/15/19	SSDI only, not in PSUs	5,048		135		90
7/15/19	All SSI, in PSUs	1,076	400	80	292	53
7/15/19	All SSI, not in PSUs	3,712		277		185
9/1/19	SSDI only, in PSUs	1,001	247	32	178	22
9/1/19	SSDI only, not in PSUs	4,079		131		87
9/1/19	All SSI, in PSUs	783	160	59	115	39
9/1/19	All SSI, not in PSUs	2,709		204		136
Total	SSDI only, in PSUs	10,788	3,922	340	3,100	227
Total	SSDI only, not in PSUs	42,634		1,338		894
Total	All SSI, in PSUs	11,345	4,096	502	2,922	334
Total	All SSI, not in PSUs	36,931		1,670		1,113
Overall total		**101,698**	**8,018**	**3,850**	**6,022**	**2,568**

Source: NBS Round 7.

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. Base sampling weights

a. Initial probability weights

We computed the initial weights for the cross-sectional 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 corresponding 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. Therefore, we ended up with two sets of initial probability weights, one each for the clustered and unclustered samples. These sets of weights both summed to the number of successful workers in the population at Round 7: 101,698.

Base weights incorporating dual sample design

To obtain estimates from the cross-sectional SWS, we had to use a "dual sample design" that combined the clustered and unclustered samples (each representing the same population) 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 by phone 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 by phone in the clustered sample, the case went to the field for additional data collection (field follow-up). Because the two samples represent the same population, we form a composite weight when combining them, multiplying the weights for one sample by λ and the weights for the other sample by 1-λ, where λ is between 0 and 1. The following section describes this in more detail.

b. 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:

(1) $$EST(Y) = \lambda Y_c + (1-\lambda) Y_u$$

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 7 data, the clustered sample accounted for 252 respondents and the unclustered sample for 76 respondents. The estimates to be combined are the proportion of the 252 in the clustered sample who are currently working and the proportion of the 76 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 in the next subsection). For the sampling variance, V(Y), the estimate is computed with the following equation:

(2) $$V(Y) = \lambda^2 V(Y_c) + (1-\lambda)^2 V(Y_u)$$

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:

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

where n_c and n_u are the sample sizes of the clustered and unclustered central office–located samples, respectively, and $deff_c$ $deff_u$ 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:

$$\lambda = V(Y_u) / \left[V(Y_c) + V(Y_u)\right] \tag{4}$$

In this case, the minimum variance is:

$$V(Y) = \left[V(Y_c) * V(Y_u)\right] / \left[V(Y_c) + V(Y_u)\right] \tag{5}$$

c. Application of composite weights to the cross-sectional SWS

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:

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

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:

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

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 to yield the base weight. 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.

2. Nonresponse adjustment

As with the RBS, we adjusted the base 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,[73] 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 similar factors as those used to locate and interview RBS members: personal characteristics of the sample member (ethnicity and age), identity of the payee with respect to the beneficiary, whether the beneficiary and the applicant for benefits lived in the same location, the number of addresses or phone numbers in the beneficiary's SSA files, the beneficiary's living situation, the beneficiary's "title" (SSI only, SSDI only, or concurrent), the beneficiary's primary disability, and geographic characteristics, including attributes of the county where the beneficiary resides. Unique to the SWS, extract was also a key factor. In Section C.2.d, we describe how the specific covariates for each of the weight adjustments varied.

[73] 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.

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 41.0 percent response rate for the cross-sectional SWS is the product of the weighted location rate and weighted completion rate among located sample members. The weighted location rate is 87.9 percent, and the weighted cooperation rate (the weighted completion rate among located sample members) is 46.4 percent. We used the weighted rates because the base weights vary substantially across the sampling strata, and the weighted rates better reflect the potential for nonresponse bias.

c. Factors related to location and cooperation

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 (base weight), weighted cooperation rate among located sample members (location-adjusted base weight), and weighted overall response rate for these factors (base weight).

Table III.7. Weighted location, cooperation, and response rates for cross-sectional SWS, by selected characteristics

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
All	8,590	6,486	87.9	3,327	46.4	41.0
Extract						
Extract 1	1,757	1,391	92.9	796	52.7	48.9
Extract 2	1,438	1,158	90.9	647	52.2	47.5
Extract 3	1,327	1,038	85.0	483	44.8	38.2
Extract 4	1,043	711	88.1	381	44.3	39.2
Extract 5	1,450	1,055	83.6	473	40.1	33.7
Extract 6	998	712	85.5	351	44.7	38.3
Extract 7	577	421	88.0	196	42.1	37.1
SSI only, SSDI only, or both SSI and SSDI						
SSI only	2,397	1,817	89.2	937	47.2	42.3
SSDI only	4,221	3,192	86.6	1,644	46.5	40.5
Both SSI and SSDI	1,972	1,477	89.6	746	45.5	40.9

Table III.7 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Constructed disability category						
Deaf	181	122	86.9	50	34.4	30.1
Cognitive disability	1,251	914	87.3	427	43.9	38.4
Mental illness	3,106	2,348	88.5	1,184	45.5	40.5
Physical disability	3,966	3,039	87.8	1,633	48.4	42.7
Unknown	86	63	84.6	33	46.6	39.1
Beneficiary's age						
18 to 29	2,078	1,514	86.5	695	41.4	36.1
30 to 39	2,075	1,545	87.8	751	43.8	38.7
40 to 49	1,864	1,386	87.7	717	46.7	41.1
50 and older	2,573	2,041	89.2	1,164	52.0	46.6
Sex						
Male	4,694	3,535	87.7	1,750	44.3	39.1
Female	3,896	2,951	88.2	1,577	49.1	43.5
Ethnicity						
Hispanic	349	254	88.1	109	38.0	33.8
Non-Hispanic	8,241	6,232	87.9	3,218	46.7	41.3
Race						
Non-Hispanic White	3,747	2,785	87.0	1,410	45.6	39.8
Non-Hispanic Black	2,490	1,940	90.0	1,040	49.6	44.7
Hispanic	349	254	88.1	109	38.0	33.8
Asian American, Pacific Island American,	73	52	80.0	22	38.3	30.9
American Indian, or Alaska Native	20	12	87.5	8	57.4	52.8
Other or unknown	1,911	1,443	87.5	738	45.8	40.3
Living situation						
Living alone	4,096	3,096	89.8	1,580	46.6	89.8
Living with others	237	173	84.6	93	45.5	84.6
Living with parents	28	17	70.3	6	26.4	70.3
In institution or unknown	4,229	3,200	86.6	1,648	46.5	40.5
Did the applicant for benefits live in the same zip code as the beneficiary?						
No	535	412	89.6	192	41.1	37.0
Yes	3,765	2,837	89.7	1,470	47.3	42.6
No information	4,290	3,237	86.4	1,665	46.4	40.3
Identity of the payee with respect to the beneficiary						
Beneficiary received payments directly	537	419	89.7	228	50.8	45.5
Payee is a family member	1,606	1,206	87.8	565	43.3	38.3
Payee is an institution	129	100	93.0	42	33.9	32.4
Other	117	82	87.7	32	36.0	32.2
Unknown	6,201	4,679	87.7	2,460	47.3	41.7

Table III.7 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Number of phone numbers in file						
Zero	553	435	88.4	250	52.3	46.8
One	1,271	921	83.2	485	48.0	40.2
Two	2,160	1,597	86.5	793	45.1	39.2
Three	2,178	1,674	90.3	875	47.4	42.9
Four	1,742	1,327	89.0	661	43.9	39.3
Five or more	686	532	90.7	263	46.6	42.5
Number of addresses in file						
Zero	547	435	89.3	249	52.2	47.1
One	1,530	1,156	87.8	599	48.0	42.5
Two	1,824	1,389	87.4	682	43.4	38.0
Three	2,227	1,678	87.3	853	45.9	40.3
Four	1,656	1,226	87.6	650	48.5	42.5
Five or more	806	602	90.8	294	44.1	40.1
Census region						
Midwest	1,840	1,356	87.8	753	49.5	43.8
Northeast	2,034	1,552	88.0	750	44.4	39.2
South	2,719	2,048	89.0	1,088	48.5	43.2
West	1,997	1,530	86.4	736	42.3	36.9
Census division						
East North Central	1,320	971	87.5	535	50.5	44.5
East South Central	535	416	90.9	228	49.4	45.2
Middle Atlantic	1,404	1,073	87.5	511	43.7	38.4
Mountain	442	333	85.6	180	44.5	38.2
New England	630	479	89.1	239	46.0	41.1
Pacific	1,555	1,197	86.8	556	41.5	36.4
South Atlantic	1,306	977	89.3	509	47.4	42.3
West North Central	520	385	88.6	218	47.4	42.3
West South Central	878	655	87.3	351	49.4	43.2
Metropolitan status of county						
Metropolitan areas with population of 1 million or more	5,123	3,938	87.6	1,980	46.0	40.6
Metropolitan areas with population of 250,000 to 999,999	2,037	1,570	89.0	813	46.6	41.6
Metropolitan areas with population of fewer than 250,000	719	506	84.7	281	49.7	42.3
Nonmetropolitan areas adjacent to large metropolitan areas	207	154	90.3	84	45.1	41.0

Table III.7 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Nonmetropolitan areas adjacent to medium or small metropolitan areas	320	213	92.4	115	46.5	43.0
Nonmetropolitan areas not adjacent to metropolitan areas	184	105	87.5	54	40.4	35.6
County with low education level						
Yes	1,144	873	87.4	433	45.5	39.9
No	7,446	5,613	88.0	2,894	46.6	41.2
County with recreation-based economy						
Yes	668	480	85.2	222	39.6	33.6
No	7,922	6,006	88.2	3,105	47.1	41.8
County with population loss						
Yes	397	244	86.1	153	58.2	50.5
No	8,193	6,242	88.0	3,174	45.9	40.6
Retirement destination county						
Yes	1,046	783	85.9	403	46.7	39.9
No	7,544	5,703	88.2	2,924	46.4	41.2
County with manufacturing-dependent economy						
Yes	640	463	85.7	247	48.4	41.9
No	7,950	6,023	88.1	3,080	46.3	41.0
County with nonspecialized-dependent economy						
Yes	6,021	4,606	88.5	2,366	47.0	41.8
No	2,569	1,880	86.7	961	45.4	39.6
County with government-dependent economy						
Yes	1,004	750	89.1	401	48.0	42.9
No	7,586	5,736	87.8	2,926	46.2	40.8
High-poverty county						
Yes	1,007	732	89.3	400	51.9	46.6
No	7,583	5,754	87.8	2,927	45.8	40.4
County with high child poverty						
Yes	1,204	900	89.2	488	50.6	45.3
No	7,386	5,586	87.7	2,839	45.9	40.4
Percentage of dwellings that are owner-occupied in county						
Less than 60.8 percent	2,805	2,145	88.5	1,080	46.1	41.0
60.8 percent to 66.2 percent	2,480	1,960	88.5	1,037	48.1	42.9
More than 66.2 percent	3,305	2,381	87.2	1,210	45.6	39.9
County racial/ethnic profile						
At least 20 percent American Indian	11	5	100.0	3	57.9	59.8
At least 90 percent non-Hispanic White	560	361	86.5	203	47.8	41.4
Plurality or majority Hispanic	849	629	87.1	307	44.0	38.6

Table III.7 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Majority but less than 90 percent non-Hispanic White	3,511	2,694	88.0	1,346	44.6	39.6
Racially/ethnically mixed, no majority group, less than 20 percent American Indian	3,291	2,520	88.1	1,321	48.7	43.0
Plurality or majority non-Hispanic Black	368	277	89.8	147	50.8	45.8
DCF earnings category, first breakdown[a]						
Gross annual DCF earnings above $30,000 in 2017 or 2018	1,966	1,469	87.4	673	40.4	35.6
Gross annual DCF earnings above $20,000 in 2017 or 2018	2,063	1,529	86.9	774	46.5	40.7
Gross annual DCF earnings above $15,000 in 2017 or 2018	1,643	1,272	89.0	685	49.1	43.8
Gross annual DCF earnings above $7,000 in 2017 or 2018	1,849	1,416	89.9	750	48.0	43.4
Gross annual DCF earnings below $7,000 in 2017 and 2018	1,069	800	85.9	445	50.2	43.1
DCF earnings category, second breakdown[a]						
Monthly DCF earnings above SGA[b] for three consecutive months in 2017 or 2018	7,355	5,563	88.1	2,845	46.3	41.0
Gross annual DCF earnings above three times SGA in 2017 or 2018	611	465	89.0	233	46.1	41.2
Gross annual DCF earnings above $0 in 2017 or 2018	301	215	82.6	128	53.5	44.1
No annual DCF earnings in 2017 or 2018	323	243	87.3	121	43.9	38.2

Source: NBS Round 7.

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

[b]Nonblind substantial gainful activity, or $1,170 in 2017, $1,180 in 2018, and $1,220 in 2019.

DCF = Disability Control File.

d. Propensity models for weight adjustments

The weight adjustments used in the cross-sectional 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 base 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: Cross-sectional SWS

Factors in location model
Main effects
EXTRACT
AGECAT (AGE CATEGORY)
SSI_SSDI (BENEFICIARY TITLE: RECIPIENT OF SSI AND/OR SSDI)
LIVING SITUATION
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
RACE
CNTYNONSP (NONSPECIFIC-DEPENDENT ECONOMY, COUNTY)
CNTYGOV (GOVERNMENT DEPENDENT ECONOMY, COUNTY)
METRO (METROPOLITAN STATUS OF COUNTY)
EARNINGS CATEGORY
Two-factor interactions
LIVING SITUATION * MOVE
RACE * MOVE

Table III.9. Cooperation logistic propensity model: SWS

Factors in cooperation model
Main effects
EXTRACT
AGECAT (AGE CATEGORY)
DISABILITY (DISABILITY CATEGORY)
EARNINGS CATEGORY
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
REPREPAYEE (IDENTITY OF PAYEE WITH RESPECT TO BENEFICIARY)
CNTYHPOV
ETHNICITY (HISPANIC OR NOT)
Two-factor interactions
EXTRACT * AGECAT

Source: NBS Round 7.

The Cox-Snell R-squared is 0.033 (0.064 when rescaled to have a maximum of 1) for the location model and 0.025 (0.033 when rescaled) for the cooperation model. 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, 64.8 percent of pairs are concordant, 34.1 percent of pairs are discordant, and the p-value for the chi-square statistic from the Hosmer-Lemeshow (H-L) goodness-of-fit test is 0.931. 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 3.57. For the cooperation model, 57.6 percent of pairs are concordant and 40.5 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.389 for the model. The cooperation adjustment from the model, which is calculated as the inverse of the cooperation propensity score, ranged from 1.34 to 5.81. The overall nonresponse adjustment (the product of the location adjustment and the cooperation adjustment) ranged from 1.52 to 6.54.

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 three levels: (1) one address on file, (2) two addresses on file, and (3) no addresses or three or more addresses on file.

- **PHONE.** Count of phone numbers in SSA files. There are five levels: Levels 1 through 4 indicate one, two, three, or four phone numbers on file, respectively, and Level 5 indicates no phone numbers or five or more phone numbers on file.
- **AGECAT.** Beneficiary's age category. There are two levels: (1) age 18 to 29 and (2) age 30 or older.
- **SSI_SSDI.** Beneficiary title. There are two levels: (1) recipient of SSDI only and (2) recipient of SSI only or of both SSI and SSDI.
- **LIVING.** Beneficiary's living situation. There are two levels: (1) beneficiary lives alone and (2) beneficiary lives with others, with parents, or in an institution or the information is unknown.
- **PDZIPSAME.** Whether the SSI beneficiary and the SSI applicant for benefits lived in the same zip code. There are three levels: (1) beneficiary and applicant lived in the same zip code, (2) beneficiary and applicant lived in different zip codes, or (3) beneficiary was a recipient of SSDI only or the information is unknown.
- **RACE**. Race of beneficiary. There are two levels: (1) non-Hispanic Black and (2) not non-Hispanic Black or race is unknown.
- **METRO**. Metropolitan status of beneficiary's county of residence. There are four levels: (1) beneficiary lived in a metropolitan area with a population of 250,000 or more; (2) beneficiary lived in a metropolitan area with a population of fewer than 250,000; (3) beneficiary lived in a nonmetropolitan area adjacent to a metropolitan area of 1 million people or more; and (4) beneficiary lived in a nonmetropolitan area adjacent to a metropolitan area of fewer than 1 million people, or beneficiary lived in a nonmetropolitan area not adjacent to a metropolitan area.
- **EARNCAT**. Earnings category from 2017 to 2018. There are four mutually exclusive levels: (1) gross annual earnings exceed SGA for three consecutive months at least once in 2017 or 2018; (2) not in Group 1, but gross annual earnings exceed three times SGA in 2017 or 2018; (3) not in Groups 1 or 2, but gross annual earnings exceed zero in 2017 or 2018; and (4) gross annual earnings are zero in both 2017 and 2018.
- **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.

The final selected model also included two interactions involving the above variables 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.
- **ETHNICITY.** Ethnicity of beneficiary. There are two levels: (1) Hispanic and (2) not Hispanic.
- **DISABILITY.** Beneficiary's disability category. There are two levels: (1) deafness and (2) hearing with other disability, or disability unknown.
- **PDZIPSAME.** Whether the SSI beneficiary and SSI applicant for benefits lived in the same zip code. There are two levels: (1) beneficiary and applicant lived in the same zip code, and(2) beneficiary and applicant lived in different zip codes, the beneficiary received SSDI only, or the information is unknown.
- **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.
- **EARNCAT.** Earnings category from 2017 to 2018. There are four mutually exclusive levels: (1) gross annual earnings exceed SGA for three consecutive months at least once in 2017 or 2018; (2) not in Group 1, but gross annual earnings exceed three times SGA in 2017 or 2018; (3) not in Groups 1 or 2, but gross annual earnings exceed zero in 2017 or 2018; and (4) gross annual earnings are zero in both 2017 and 2018.
- **CNTYHPOV.** County with high levels of poverty. There are two levels: (1) county where 20 percent or more of its residents were poor, based on the American Community Survey's five-year estimates for 2008 to 2012, and (2) county does not have this attribute.

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.

3. Post-stratification and trimming

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

Table III.10. Design effects attributable to unequal weights before and after trimming, within trimming classes in the cross-sectional SWS

Extract	Sampling stratum	Number of cases trimmed	Design effect attributable to unequal weights	
			Before trimming	After trimming
1	SSDI only	1	1.19	1.19
1	SSI	2	1.24	1.23
2	SSDI only	0	1.25	1.25
2	SSI	0	1.19	1.19
3	SSDI only	0	1.24	1.24
3	SSI	0	1.24	1.24
4	SSDI only	3	1.41 (maximum)	1.34
4	SSI	0	1.17	1.17
5	SSDI only	0	1.25	1.25
5	SSI	0	1.15	1.15
6	SSDI only	0	1.37	1.37 (maximum)
6	SSI	0	1.16	1.16
7	SSDI only	1	1.34	1.16
7	SSI	0	1.19	1.19

Source: NBS Round 7.

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

After the nonresponse adjustment and trimming, we poststratified the weights to the population totals for four variables: (1) extract; (2) beneficiary title (SSI only, SSDI only, and both SSI and SSDI); (3) four age categories (18 to 29, 30 to 39, 40 to 49, and 50 or over); and DCF earnings (five categories derived from DCF earnings in 2017 and 2018—the categorization of earnings listed under "first breakdown" in Table III.7). We found no extreme weights after poststratification.

D. Longitudinal SWS

The Round 7 longitudinal sample consists of the Round 6 cross-sectional SWS respondents who indicated that they were working at the time of the Round 6 interview. Table III.11 presents the final sample sizes for the longitudinal SWS. This table shows a final sample of 2,404 cases from the Round 6 clustered sample and 1,308 from the Round 6 unclustered sample, for a total of 3,712 sample cases, of which 216 were selected for both the clustered and unclustered samples in Round 6. We do not know what proportion of the 89,636 successful workers in Round 6 were working at the time of the Round 6 interview, but we have an estimate based on our responding sample, which is shown in Table III.11 (65,871), of which 64,225 were eligible. However, after we processed an updated extract from Round 6, we found that there was a total of 288,576 successful workers, of which 265,514 were eligible. We poststratified the Round 6 weights to

this new total; however, we still need to recalculate the longitudinal weights to determine an estimated size of the eligible longitudinal population.[74]

For the sake of brevity, Table III.11 does not break out results by Round 6 extract or by whether the unclustered case was in a PSU in Round 6, as these stratification variables are not analytically useful. Moreover, data collection for all Round 6 extraction dates occurred simultaneously in Round 7. Theoretically, the same follow-up rules for the clustered and unclustered samples were used in Round 7 as were used in Round 6;[75] however, we followed up clustered cases in the field if they also happened to be sampled for the Round 7 RBS or were in the clustered sample for the Round 7 cross-sectional SWS.

Table III.11. Estimated survey population and sample sizes, by beneficiary title strata in the longitudinal SWS

Stratum[a]	Weighted total	Clustered sample in Round 6	Unclustered sample in Round 6	Total sample in Round 6
SSDI only	33,675.7	1,180	683	1,863
All SSI	32,195.1	1,224	625	1,849
Total	65,870.8	2,404	1,308	3,712

Source: NBS Round 7.

[a]These stratification variables are defined based on the sample member's situation on June 30, 2016. For some longitudinal sample members who were still SSI or SSDI beneficiaries on June 30, 2018, their beneficiary title had changed.

1. Base sampling weights

a. Initial probability weights

We used the final weights for the Round 6 SWS as the "initial probability weights" for the Round 7 longitudinal SWS. The 3,712 cases in the longitudinal sample included 108 duplicates (216 sample cases) across the clustered and unclustered samples. For an additional 20 duplicates (40 sample cases), the Round 6 completed interviews in the clustered sample were obtained due to field efforts. Therefore, the 20 cases in the unclustered sample were represented by the clustered sample, and the 20 cases in the unclustered sample had their Round 6 cross-sectional weight set to zero. For this reason, these 20 cases were not included among the 3,712 longitudinal sample cases.

[74] After we conducted a final extract of Round 6 earnings data in November 2020, we determined that the estimated number of eligible successful workers in Round 6 was actually 265,514; the discrepancy was due to a lag in recording earnings in SSA administrative data for many successful workers. Since it takes three years for this lag to dissipate, we will also need to redo the Round 7 longitudinal weights in 2022 to account for this new total and obtain a new estimate of successful workers who were eligible for the longitudinal population..

[75] In practice, to save resources, longitudinal SWS cases that should have been sent to the field in Round 7 (clustered in Round 6) were often not.

b. Base weights incorporating dual sample design

The Round 6 cross-sectional final weights already accounted for the dual sample design, so it was not necessary to recreate the composite weights. However, because of different data collection dispositions in Round 7 than in Round 6, we needed to account for the different field follow-up rules between the clustered and unclustered samples (rules that were supposed to be consistent between the two rounds).[76]

In particular, for sample members the population that did not need field operations to resolve in Round 7 (cases completed by phone), we used the weights as they were, regardless of clustered or unclustered status, and regardless of whether completed by phone or field in Round 6 for clustered. However, if sample members came from the population that needed field operations in Round 7 to resolve (cases not completed by phone), we estimated the size of this population by summing the weights of the Round 7 field-resolution cases (cases not able to be completed by phone). We then set the weights of the Round 6 unclustered sample in this population to zero, and we ratio-adjusted the weight of the clustered sample to match this estimated total. There were 42 such unclustered sample cases with weights set to zero. Therefore, the number of longitudinal SWS sample members with nonzero base weights was 3,670 (3,712 − 42).

2. Nonresponse adjustment

As indicated earlier, when calculating the nonresponse adjustments, we separated the Round 7 longitudinal SWS into two groups, depending on whether the sample member was still an SSI or SSDI beneficiary as of June 30, 2018. We did this for two reasons: (1) there are likely important differences between the longitudinal sample members who were or were not part of the Round 7 beneficiary frame, and (2) for members who were part of the Round 7 beneficiary frame, we could use auxiliary variables from that frame. However, for sample members who were not part of that frame, we could only use Round 7 geographically based information. All other covariates had to come from the Round 6 frame.

For both groups, we adjusted the base weights in two stages for (1) sample members who could not be located and (2) sample members who were located but refused to respond. The group in the Round 7 beneficiary frame consisted of 3,182 of 3,712 longitudinal sample members (or 3,147 of 3,670 with positive base weights). We used weighted logistic propensity models to calculate the location adjustment for all members of this group and the cooperation adjustments for located members of this group. But for those who were not in the Round 7 beneficiary frame (530 of 3,712 sample members, or 523 with positive base weights), we calculated the adjustments using simple weighting classes due to the small sample size and more limited information available.

[76] We assumed that all clustered longitudinal cases would use the same field follow-up rules in Round 7, even though in practice (in all but three cases) we did not use field follow-up for clustered cases if they were not also sampled in Round 7 as part of the RBS or the clustered cross-sectional SWS.

For the 3,147 longitudinal sample cases with positive base weights that were part of the Round 7 beneficiary frame, we fit the models in the same way as the adjustment models for the RBS (Section B.2 of this chapter) and cross-sectional SWS (Section C.2 of this chapter). For the remaining 523 longitudinal cases with positive base weights that were not part of the Round 7 beneficiary frame, we fit cross-tabulations and stepwise logistic regression models to identify factors to use in the weighting classes.

The main factors or attributes that affected our ability to locate and interview longitudinal SWS members of both types included similar factors to those used to locate and interview RBS and cross-sectional SWS members: personal characteristics of the sample member (race and age); whether the beneficiary and applicant for benefits lived in the same location; the number of addresses or phone numbers in the beneficiary's SSA files; the beneficiary's living situation; the beneficiary's "title" (SSI only, SSDI only, or concurrent); the beneficiary's primary disability; and geographic characteristics, including attributes of the county where the beneficiary lives. As with the cross-sectional SWS, extract was also a key factor. For the longitudinal successful workers who were not part of the Round 7 beneficiary frame, variables that were only available from the Round 7 frame had to come from the Round 6 frame. In Section D.2.d, we describe how the specific covariates for each set of weight adjustment varied.

a. Coding of survey dispositions

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

b. Response rates

The 54.5 percent response rate for the longitudinal SWS is the product of the weighted location rate and weighted cooperation rate among located sample members. The weighted location rate is 89.1 percent, and the weighted cooperation rate (the weighted completion rate among located members) is 60.1 percent. Analogous to the RBS and cross-sectional SWS, we used the weighted rates because the base weights vary greatly across the sampling strata, and the weighted rates better reflect the potential for nonresponse bias.

c. Factors related to location and cooperation

Table III.12 shows selected factors associated with locating a sample member and the factors associated with the response among located sample members for those who were part of the Round 7 frame. Table III.13 shows these factors for sample members who were not part of the frame. The tables include unweighted counts of all sample members, counts of located sample members, and counts of sample members who had a completed interview or were deemed ineligible. The tables also include the weighted location rate (base weight), weighted cooperation rate among located sample members (location-adjusted base weight), and weighted overall response rate for these factors (base weight). In both tables, the first row provides the overall counts and response rates for reference.

Table III.12. Weighted location, cooperation, and response rates for longitudinal SWS, by selected characteristics, among those in Round 7 beneficiary frame

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
All longitudinal successful workers	3,670	3,313	89.1	2,114	60.9	54.5
Longitudinal successful workers in Round 7 beneficiary frame	3,147	2,859	89.9	1,868	62.3	56.2
Extract						
Extract 1	675	612	89.0	428	66.8	59.4
Extract 2	502	462	90.6	298	62.2	56.5
Extract 3	514	457	88.4	266	54.9	48.9
Extract 4	412	368	86.9	225	58.5	50.8
Extract 5	409	377	91.6	249	63.1	57.8
Extract 6	301	276	91.9	183	64.2	58.9
Extract 7	334	307	92.9	219	70.4	65.5
SSI only, SSDI only, or both SSI and SSDI						
SSI only	1,053	960	90.7	622	61.7	56.2
SSDI only	1,563	1,415	89.2	949	64.8	57.8
Both SSI and SSDI	531	484	90.5	297	56.3	51.5
Constructed disability category						
Deaf	53	48	86.1	20	48.0	40.9
Cognitive disability	387	355	89.0	214	56.9	50.9
Mental illness	1,084	981	90.8	628	61.8	56.4
Physical disability	1,573	1,432	89.9	974	64.8	58.3
Unknown	50	43	84.6	32	67.0	57.4
Beneficiary's age						
18 to 29	726	659	90.2	389	55.7	50.4
30 to 39	652	585	89.2	352	56.3	50.6
40 to 49	673	604	89.9	400	65.7	59.0
50 and older	1,096	1,011	90.3	727	70.4	63.6
Sex						
Male	1,608	1,470	90.0	943	60.2	54.5
Female	1,539	1,389	89.8	925	64.8	58.2
Ethnicity						
Hispanic	168	154	91.2	94	59.9	55.4
Non-Hispanic	2,979	2,705	89.9	1,774	62.4	56.2
Race						
Non-Hispanic White	1,293	1,170	89.1	761	62.4	55.6
Non-Hispanic Black	811	753	92.7	529	67.7	62.8
Hispanic	168	154	91.2	94	59.9	55.4
Asian American or Pacific Island American	34	30	93.0	20	58.2	53.9
American Indian or Alaska Native	6	6	100.0	4	72.2	71.3
Other or unknown	835	746	88.5	460	57.7	51.2
Living situation						
Living alone	1,503	1,369	90.4	879	60.6	55.0

Table III.12 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Living with others	67	63	95.8	33	46.0	45.6
Living with parents	8	6	76.0	4	61.5	47.7
In institution or unknown	1,569	1,421	89.3	952	64.7	57.7
Did the applicant for benefits live in the same zip code as the beneficiary?						
No	147	135	95.1	82	60.1	57.2
Yes	1,409	1,282	89.9	821	59.8	54.2
No information	1,591	1,442	89.4	965	64.6	57.7
Identity of the payee with respect to the beneficiary						
Beneficiary received payments directly	153	141	91.0	81	49.7	45.2
Payee is a family member	494	447	90.4	278	59.4	53.7
Payee is an institution	40	39	97.1	23	59.2	57.5
Other	39	38	77.9	23	62.1	48.4
Unknown	2,421	2,194	89.9	1,463	64.2	57.7
Number of phone numbers in file						
Zero	530	502	95.0	358	68.0	64.7
One	393	362	91.8	226	61.4	56.4
Two	740	672	90.0	434	60.3	54.5
Three	718	631	86.3	420	62.5	54.2
Four	543	490	88.5	306	60.2	53.4
Five or more	223	202	89.7	124	61.1	54.4
Number of addresses in file						
Zero	192	187	96.5	135	69.2	66.7
One	640	577	89.2	397	65.3	59.4
Two	696	639	90.6	400	58.0	52.9
Three	838	745	87.5	468	60.0	52.7
Four	538	488	89.9	320	64.7	58.1
Five or more	243	223	91.7	148	62.7	57.2
Census region						
Midwest	674	612	89.3	415	64.5	57.9
Northeast	826	772	93.3	488	58.3	54.4
South	923	817	87.8	533	62.5	55.0
West	724	658	89.9	432	64.1	57.8
Census division						
East North Central	475	435	91.3	302	66.3	60.8
East South Central	161	151	91.8	100	61.4	56.6
Middle Atlantic	561	520	92.1	336	58.8	54.2
Mountain	157	146	91.1	96	63.4	58.0
New England	265	252	96.3	152	57.0	54.9
Pacific	567	512	89.5	336	64.4	57.7
South Atlantic	482	422	87.1	272	62.9	55.1
West North Central	199	177	85.1	113	60.7	52.1
West South Central	280	244	86.3	161	62.4	54.0

Table III.12 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Metropolitan status of county						
Metropolitan area with population of 1 million or more	1,934	1,770	91.2	1,154	62.7	57.3
Metropolitan area with population of 250,000 to 999,999	730	658	89.5	432	63.2	56.6
Metropolitan area with population of fewer than 250,000	218	188	82.2	123	60.4	50.0
Nonmetropolitan area adjacent to large metropolitan areas	69	59	88.7	32	50.1	44.3
Nonmetropolitan area adjacent to medium or small metropolitan areas	109	102	95.0	75	66.4	63.1
Nonmetropolitan area not adjacent to metropolitan areas	87	82	91.6	52	59.8	54.3
County with low education level						
Yes	442	395	88.1	255	63.7	56.3
No	2,705	2,464	90.2	1,613	62.1	56.2
County with recreation-based economy						
Yes	215	192	89.5	118	59.9	53.7
No	2,932	2,667	90.0	1,750	62.5	56.4
County with population loss						
Yes	137	127	92.1	83	62.3	57.5
No	3,010	2,732	89.8	1,785	62.3	56.1
Retirement destination county						
Yes	323	284	87.5	169	55.7	48.9
No	2,824	2,575	90.3	1,699	63.2	57.2
County with manufacturing-dependent economy						
Yes	187	168	85.4	115	65.1	55.8
No	2,960	2,691	90.4	1,753	62.0	56.2
County with nonspecialized-dependent economy						
Yes	2,275	2,073	90.7	1,350	61.8	56.2
No	872	786	88.4	518	63.2	56.1
County with government-dependent economy						
Yes	374	337	88.5	223	62.5	55.7
No	2,773	2,522	90.2	1,645	62.3	56.2
High-poverty county						
Yes	395	363	91.1	245	66.6	60.6
No	2,752	2,496	89.8	1,623	61.7	55.6
County with high level of child poverty						
Yes	461	424	90.1	288	65.7	59.2
No	2,686	2,435	89.9	1,580	61.8	55.7
Percentage of dwellings that are owner occupied in county						
Less than 60.8 percent	1,112	1,010	90.3	666	63.7	57.7
60.8 percent to 66.2 percent	934	850	89.1	549	61.6	55.1
More than 66.2 percent	1,101	999	90.2	653	61.8	55.8

Table III.12 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
County racial/ethnic profile						
At least 20 percent American Indian	8	8	100.0	6	68.4	68.2
At least 90 percent non-Hispanic White	214	196	90.6	125	58.4	53.2
Plurality or majority Hispanic	317	284	88.5	181	63.6	56.2
Majority but less than 90 percent non-Hispanic White	1,155	1,046	89.4	680	61.8	55.3
Racially/ethnically mixed, no majority group, less than 20 percent American Indian	1,304	1,188	90.5	784	63.4	57.6
Plurality or majority non-Hispanic Black	149	137	91.5	92	63.9	58.5
Beneficiary's DCF earnings category[a]						
Gross annual DCF earnings above $30,000 in 2017 or 2018	646	593	91.6	392	62.8	57.2
Gross annual DCF earnings above $20,000 in 2017 or 2018	710	638	89.5	413	61.9	55.7
Gross annual DCF earnings above $15,000 in 2017 or 2018	544	499	90.6	339	62.5	56.8
Gross annual DCF earnings above $7,000 in 2017 or 2018	703	642	91.3	421	63.1	57.9
Gross annual DCF earnings below $7,000 in 2017 and 2018	544	487	86.2	303	60.9	52.7

Source: NBS Round 7.

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

Table III.13. Weighted location, cooperation, and response rates for longitudinal SWS, by selected characteristics, among those not in Round 7 beneficiary frame

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
All longitudinal successful workers	3,670	3,313	89.1	2,114	60.9	54.5
Longitudinal successful workers not in Round 7 beneficiary frame	523	454	84.8	246	53.1	45.3
Extract						
Extract 1	123	106	83.0	58	57.0	47.3
Extract 2	79	67	83.4	36	49.9	42.0
Extract 3	83	72	86.2	36	43.6	37.5
Extract 4	61	53	88.4	26	46.4	41.6
Extract 5	57	47	80.8	26	60.3	49.3
Extract 6	44	39	82.4	21	60.1	49.8
Extract 7	76	70	91.9	43	61.7	56.8
SSI only, SSDI only, or both SSI and SSDI in Round 6						
SSI only	73	63	85.5	29	43.1	36.8
SSDI only	396	348	86.4	198	56.5	49.0
Both SSI and SSDI	54	43	71.8	19	40.4	28.9
Constructed disability category in Round 6						
Cognitive disability	35	30	84.9	11	33.1	27.6
Mental illness	184	160	84.1	95	57.5	48.8
Physical disability, including deafness, or unknown	304	264	85.3	140	52.5	45.1
Beneficiary's Round 7 age						
18 to 29	63	55	83.1	23	44.4	36.7
30 to 39	110	89	83.0	39	43.2	36.0
40 to 49	132	114	83.9	60	52.1	44.0
50 and older	218	196	87.3	124	64.1	56.2
Sex						
Male	261	220	82.2	120	54.3	44.7
Female	262	234	87.8	126	51.9	46.0
Race in Round 6						
Non-Hispanic White	255	211	81.5	115	53.5	43.7
Non-Hispanic Black	156	145	93.2	79	52.5	49.4
Hispanic, other races, or unknown	112	98	83.1	52	53.0	44.6
Living situation in Round 6						
Living alone	118	100	80.9	45	41.3	33.4
Living with others, parents, in institution or unknown	405	354	85.9	201	56.5	48.6
Did the applicant for benefits live in the same zip code as the beneficiary in Round 6?						
Yes	109	91	80.0	41	41.4	33.2
No, or no information	414	363	86.0	205	56.1	48.3

Table III.13 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Identity of the payee with respect to the beneficiary in Round 6						
Beneficiary received payments directly	33	28	81.2	18	64.2	53.0
Payee is a family member	59	50	83.3	24	48.0	39.9
Payee is an institution, other, or unknown	431	376	85.2	204	53.4	45.6
Number of phone numbers in file in Round 7						
Zero	70	65	94.1	39	61.4	57.7
One	74	66	83.4	34	49.0	41.4
Two	113	96	82.2	51	50.0	41.1
Three	130	109	81.3	56	53.9	43.9
Four	98	86	88.2	50	55.2	49.5
Five or more	38	32	85.0	16	49.8	42.5
Number of addresses in file in Round 7						
One	84	69	77.9	45	60.0	47.4
Two	114	97	82.7	52	48.1	40.0
Three	158	136	85.7	57	46.9	40.0
Four	106	95	86.0	53	55.0	48.0
Five or more	61	57	92.9	39	63.1	58.8
Census region in Round 6						
Midwest	105	89	84.3	52	57.9	49.4
Northeast	114	97	86.0	48	48.9	41.8
South	184	161	85.6	85	47.7	41.3
West	120	107	83.1	61	60.1	50.4
Census division in Round 6						
East North Central	67	58	86.2	36	58.1	50.4
East South Central	38	32	84.0	15	43.1	36.8
Middle Atlantic	77	63	84.3	31	49.7	41.4
Mountain	36	28	71.8	14	57.4	41.0
New England	37	34	89.7	17	47.3	42.8
Pacific	84	79	91.4	47	61.6	57.3
South Atlantic	104	94	85.9	48	42.9	36.9
West North Central	38	31	81.2	16	57.6	47.8
West South Central	42	35	86.9	22	66.2	58.2
Metropolitan status of county						
Metropolitan area with population of 1 million or more	327	285	86.4	151	53.5	46.6
Metropolitan area with population of 250,000 to 999,999	124	110	83.4	62	58.1	48.5
Metropolitan area with population of fewer than 250,000	31	26	85.4	13	46.5	39.5
Nonmetropolitan area	41	33	80.5	20	46.0	37.5

Table III.13 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
County with low education level						
Yes	61	52	87.4	29	55.7	50.1
No	462	402	84.5	217	52.9	44.8
County with recreation-based economy						
Yes	33	30	87.5	21	72.7	64.1
No	490	424	84.5	225	51.1	43.4
County with population loss						
Yes	23	22	97.1	13	59.7	58.3
No	500	432	84.3	233	52.8	44.8
Retirement destination county						
Yes	74	65	83.5	33	49.9	41.7
No	449	389	85.0	213	53.7	46.0
County with manufacturing-dependent economy						
Yes	25	23	92.0	13	41.0	38.1
No	498	431	84.3	233	54.1	45.8
County with nonspecialized-dependent economy						
Yes	407	352	84.1	190	52.9	44.7
No	116	102	86.6	56	53.8	47.0
County with government-dependent economy						
Yes	44	39	84.8	17	44.9	38.1
No	479	415	84.8	229	54.0	46.1
High-poverty county						
Yes	49	41	85.4	22	42.3	37.3
No	474	413	84.7	224	54.1	46.1
County with high level of child poverty						
Yes	76	67	87.2	33	43.1	38.1
No	447	387	84.4	213	54.7	46.5
Percentage of dwellings that are owner occupied in county						
Less than 60.8 percent	157	137	86.5	71	52.4	45.9
60.8 percent to 66.2 percent	171	150	83.7	78	55.1	45.9
More than 66.2 percent	195	167	84.5	97	52.2	44.5
County racial/ethnic profile						
Majority non-Hispanic White	258	217	81.9	120	53.6	44.0
Racially/ethnically mixed, no majority group, less than 20 percent American Indian	203	182	87.7	98	51.5	45.7
Other racial mixes	62	55	91.1	28	56.1	51.5
Beneficiary's DCF earnings category[a]						
Gross annual DCF earnings above $30,000 in 2017 or 2018	134	110	79.0	57	51.2	40.7
Gross annual DCF earnings above $20,000 in 2017 or 2018	141	127	88.0	73	55.5	48.9

Table III.13 (*continued*)

	Sample	Located sample		Response among located sample		Overall respondents
	Count	Count	Location rate	Count	Cooperation rate	Response rate
Gross annual DCF earnings above $15,000 in 2017 or 2018	88	82	93.8	43	51.9	48.8
Gross annual DCF earnings above $7,000 in 2017 or 2018	105	83	75.6	46	53.4	40.7
Gross annual DCF earnings below $7,000 in 2017 and 2018	55	52	96.3	27	52.1	51.4

Source: NBS Round 7.

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

d. Propensity models for weight adjustments among longitudinal SWS cases in Round 7 beneficiary frame

The weight adjustments used in the longitudinal SWS among sample cases in the Round 7 beneficiary frame were based on predicted propensities from a logistic regression model. The model-fitting process was similar to that used in the RBS and cross-sectional SWS. We identified candidate interactions using CHAID, identified variables to investigate further using the STEPWISE procedure in SAS, and then created parsimonious models using SURVEYLOGISTIC in SAS and the RLOGIST procedure in SUDAAN. As stated earlier, we calculated the adjustments by taking the inverse of the predicted location and cooperation propensities. Note that we defined these variables in terms of the beneficiary's status in Round 7. For example, the beneficiary title is based on whether the person was receiving SSI and/or SSDI benefits as of June 30, 2018, not as of June 30, 2016. Thus, their beneficiary title in Round 7 may not be consistent with their stratum assignments in Round 6.

Tables III.14 and III.15 summarize the variables included in the final location and cooperation propensity models. (Appendix D describes how we collapsed the levels for each model.)

Table III.14. Location logistic propensity model: Longitudinal SWS in Round 7 beneficiary frame

Factors in location model
Main effects
EXTRACT
AGECAT (AGE CATEGORY)
REGION (CENSUS REGION)
SSI_SSDI (BENEFICIARY TITLE: RECIPIENT OF SSDI, SSI, OR BOTH)
PDZIPSAME (WHETHER APPLICANT FOR BENEFITS LIVES IN SAME ZIP CODE AS BENEFICIARY)
PHONE (CATEGORIZED COUNT OF PHONE NUMBERS IN SSA FILES)
RACE
METRO (METROPOLITAN STATUS OF COUNTY)
CNTYGOV (GOVERNMENT-DEPENDENT ECONOMY, COUNTY)
CNTYNOFUEL (CATEGORIZED PERCENTAGE OF HOUSEHOLDS THAT DO NOT USE FUEL)
Two-factor interactions
AGECAT * CNTYGOV

Source: NBS Round 7.

Table III.15. Cooperation logistic propensity model: Longitudinal SWS in Round 7 beneficiary frame

Factors in cooperation model
Main effects
EXTRACT
AGECAT (AGE CATEGORY)
SSI_SSDI (BENEFICIARY TITLE: RECIPIENT OF SSDI, SSI, OR BOTH)
MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)
RACE
REGION (CENSUS REGION) or DIVISION (CENSUS DIVISION)
LIVING SITUATION
CNTYRET (COUNTY WITH HIGH PERCENTAGE OF RETIREES)
Two-factor interactions
BENEFICIARY TITLE (BENEFICIARY OF SSDI, SSI, OR BOTH) * EXTRACT
BENEFICIARY TITLE (BENEFICIARY OF SSDI, SSI, OR BOTH) * MOVE (CATEGORIZED COUNT OF ADDRESSES IN SSA FILES)

Source: NBS Round 7.

The Cox-Snell R-squared is 0.036 (0.075 when rescaled to have a maximum of 1) for the location model and 0.046 (0.063 when rescaled) for the cooperation model. 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, 62.3 percent of pairs are concordant, 36.6 percent of pairs are discordant, and the p-value for the chi-square statistic from the Hosmer-Lemeshow (H-L) goodness-of-fit test is 0.567. These values indicate a reasonably good fit of the model to the data. The location adjustments from the model, calculated as the inverse of the location propensity score, ranged from 1.01 to 2.00. For the cooperation model, 60.6 percent of pairs are concordant and 38.6 percent of pairs are discordant. The p-value for the chi-squared statistic for the H-L goodness-of-fit test is 0.944 for the model. The cooperation adjustments from the model, which is calculated as the inverse of the cooperation propensity score, ranged from 1.16 to 4.35. The overall nonresponse adjustments (the product of the location adjustment and the cooperation adjustment) ranged from 1.18 to 6.17.

Among the variables used in the location and cooperation models shown in Tables III.14 and III.15, the number of levels used in the models is often fewer than the number of levels in Table III.14; the levels collapsed for the models are described following the tables. The factors used in the location model included the following:

- **EXTRACT.** There are three levels: (1) Extract 5, (2) Extract 6, and (3) Extracts 1 through 4 and 7.
- **PHONE.** Count of phone numbers in SSA files. There are six levels: Levels 1 through 5 indicate zero, one, two, three, or four phone numbers on file, respectively, and Level 6 indicates five or more phone numbers on file.
- **REGION.** Geographic region of beneficiary's place of residence, based on U.S. census regions. There are four levels: (1) West, (2) South, (3) Midwest, and (4) Northeast.
- **AGECAT.** Beneficiary's age category. There are four levels: (1) ages 18 to 29, (2) ages 30 to 39, (3) ages 40 to 49, and (4) ages 50 or older.
- **RACE.** Race of beneficiary. There are two levels: (1) non-Hispanic Black and (2) not non-Hispanic Black or race not known.
- **SSI_SSDI.** Beneficiary title. There are two levels: (1) recipient of SSDI only and (2) recipient of SSI only or of both SSI and SSDI.
- **PDZIPSAME.** Whether the SSI beneficiary and the SSI applicant for benefits live in the same zip code. There are two levels: (1) the beneficiary and applicant live in the same zip code and (2) the beneficiary and applicant live in different zip codes, the beneficiary is a recipient of SSDI only, or the information is unknown.
- **METRO.** Metropolitan status of beneficiary's county of residence. There are three levels: (1) the beneficiary lives in a metropolitan area with a population between 250,000 and 1,000,000; (2) the beneficiary lives in a metropolitan area with a population of fewer than

250,000; and (3) the beneficiary lives in a metropolitan area with a population over 1,000,000 or the beneficiary lives in a nonmetropolitan area.

- **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 are derived from the federal and state government, or 9 percent or more jobs are in the federal or state government during 2010–2012, and (2) a county without this attribute.
- **CNTYNOFUEL.** Categorized percentage of occupied housing units in the county that do not use fuel. There are three levels: (1) the county's percentage of housing units that do not use fuel is less than 0.4 percent; (2) the county's percentage of housing units that do not use fuel is between 0.4 and 0.6 percent; and (3) the county's percentage of housing units that do not use fuel exceeds 0.6 percent.

The final selected model also included two interactions involving the above variables for locating sample members. Table III.14 provides the main effects, using the variable names listed above. Appendix D provides the parameter estimates and their standard errors. The factors used in the cooperation model included the following:

- **EXTRACT.** There are four levels: (1) Extract 1; (2) Extract 3; (3) Extract 7; and (4) Extracts 2, 4, 5, and 6.
- **SSI_SSDI.** Beneficiary title. There are two levels: (1) recipient of both SSI and SSDI and (2) recipient of SSDI only or SSI only.
- **MOVE.** Count of addresses in SSA files. There are four levels: (1) one address on file, (2)–(3) two or three addresses on file, and (4) four or more addresses on file.
- **AGECAT.** Beneficiary's age category. There are four levels: (1) ages 18 to 29, (2) ages 30 to 39, (3) ages 40 to 49, and (4) ages 50 or older.
- **RACE.** Race of beneficiary. There are two levels: (1) non-Hispanic Black and (2) not non-Hispanic Black or race not known.
- **LIVING.** Beneficiary's living situation. There are two levels: (1) beneficiary lives with others, and (2) beneficiary lives alone, with parents, or in an institution or the information is unknown.
- **REGION or DIVISION.** Geographic region or division of beneficiary's place of residence, based on U.S. census regions or divisions. There are three levels: (1) South, (2) West, (3) East North Central division of Midwest, and (4) West North Central division of Midwest or Northeast.
- **CNTYRET.** Retirement destination county. There are two levels: (1) the number of residents ages 60 and older grew by 15 percent or more between the 2000 and 2010 censuses due to net migration, and (2) the county does not have this attribute.

The model also included a single interaction between two of these variables for responding sample members, as noted in Table III.15. Table III.15 describes the main effects using the variable names. Appendix D provides an expanded form of Table III.15, with parameter estimates and their standard errors.

Because there were only 523 longitudinal cases that were not part of the Round 7 beneficiary frame, and only 246 completed interviews, the options for creating nonresponse adjustments for this group were limited. We used stepwise regression and cross-tabulations to determine which variables were most closely related to location and which were related to cooperation. For the location adjustment, we created four weighting classes based on the strata derived from the beneficiary title, as defined in Round 6 (SSDI only and SSI) and race (non-Hispanic White or not). The adjustments ranged from 1.08 to 1.41. For the cooperation adjustment, we created eight weighting classes based on the same Round 6 strata (SSDI only and SSI) and the four age categories (18 to 29, 30 to 39, 40 to 49, and 50 or over). These adjustments ranged from1.34 to 3.22, and the total adjustments (the product of the location and cooperation adjustments) ranged from 1.69 to 4.02.

3. Post-stratification and trimming

The adjusted weight for each sample case is the product of the base weight and the adjustment factors, trimmed to ensure that the impact of outlier weights is minimized. We performed the trimming across the two groups (both on and off the Round 7 beneficiary frame) together.

We created 14 trimming classes for each model based on the original strata from Round 6, which were in turn based on (1) the two beneficiary title levels (SSDI only and SSI), and (2) the seven extracts. We trimmed seven weights within these 14 trimming classes. Table III.16 shows the number of weights trimmed and the design effects attributable to unequal weighting before and after trimming for each class, before poststratification.

Table III.16. Design effects attributable to unequal weights before and after trimming, within trimming classes in the longitudinal SWS

Extract	Sampling stratum	Number of cases trimmed	Design effect attributable to unequal weights	
			Before trimming	After trimming
1	SSDI only	0	1.40	1.40
1	SSI	2	1.56	1.55 (maximum)
2	SSDI only	1	1.60 (maximum)	1.52
2	SSI	0	1.37	1.37
3	SSDI only	0	1.41	1.41
3	SSI	1	1.38	1.38
4	SSDI only	1	1.48	1.43
4	SSI	0	1.27	1.27
5	SSDI only	1	1.58	1.48
5	SSI	0	1.30	1.30
6	SSDI only	0	1.36	1.36
6	SSI	0	1.27	1.27
7	SSDI only	0	1.29	1.29
7	SSI	1	1.38	1.35

Source: NBS Round 7.

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

After the nonresponse adjustment and trimming, we post-stratified the weights to marginal population totals for four variables: (1) extract; (2) beneficiary title as defined in Round 6 (SSI only, SSDI only, and both SSI and SSDI); (3) four age categories (18 to 29, 30 to 39, 40 to 49, and 50 or over); and (4) DCF earnings categories in Round 6 (five categories derived from DCF earnings in 2015 and 2016—the same categories used for the SWS nonresponse models in Round 6). The actual population totals were not available, so we used the estimated totals by summing the base weights for each level of these variables. We found no extreme weights after poststratification.

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 three separate samples (the RBS, the cross-sectional SWS, and the longitudinal SWS), with duplicates occurring across and within samples. For the purpose of imputation processing, we grouped all three samples together as a single set of records requiring imputation, with duplicates removed, resulting in 8,824 records total. Where appropriate, we used the sample that the record belonged to as a covariate in the imputation.

In most cases, we used 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.[77]

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.[78] To develop the imputation classes, we used information from both the interview and SSA administrative data

[77] 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 20 percent gave only general categories of income. Detailed levels of missingness are given for all imputed variables in later sections of this chapter.

[78] 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).

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 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 actual number of imputed responses to the follow-up question to be greater than the number of nonlegitimate 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).[79] 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, a nonzero or nonmissing flag means we imputed one or more of the component variables in the constructed variable. All variables that include any 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.

[79] 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 Rounds 5 or 6 but did not in Round 7.

Table IV.1. Race and ethnicity imputations

Variable name	Description	Imputation method	Number missing	Number eligible	Percentage imputed
L1_i	Hispanic/Latino ethnic origins	5 imputations from SSA's administrative data, 28 longitudinal imputation, 241 imputations from hot deck	274	8,824	2.73
C_Race_i	Race	282 imputations from SSA's administrative data, 37longitudinal imputation, 336 imputations from hot deck	655	8,824	3.81

Source: NBS Round 7.

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 Social Security Administration's (SSA's) administrative data so indicated. Because this round included a longitudinal component, we expected to use a larger number of longitudinal imputations than in prior rounds. Indeed, there were 28 instances in which a sample member—a unit respondent in Round 7 and in at least one of Rounds 5 or 6—did not respond to L1 in Round 7 but did respond to it in Rounds 5 or 6, so we used his or her latest available response from the prior rounds. For respondents who still had missing data, we imputed the Hispanic indicator by using a hot deck imputation. 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).[80] For those without Hispanic surnames, we defined imputation 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. A process that we automated 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 (2018−2019), 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

[80] This methodology is consistent with the procedure followed in Round 6, which was a change from earlier rounds. In Rounds 1 to 5, we logically assigned "Hispanic" if an individual had a Hispanic surname.

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 were 37 instances in which a sample member—a unit respondent in Round 7 and in at least one of Rounds 5 and 6—did not respond to L2 in Round 7, but the member did respond to it in Round 5 or 6, so we used his or her latest available response from the prior round. 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 28 such cases, 16 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 did not map exactly to those used in Round 5.[81] 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

[81] For a detailed exposition of the disability codes, see the User's Guide (Callahan, et al. 2021).

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	358 hot deck[a]	358	7,145	5.01
C_MainConColDiagGrp_i	Main condition diagnosis group collapsed	358 constructed from imputed variablesa	358	7,145	5.01
C_MainConBodyGroup_i	Main condition body group	29 hot deck, 329 constructed from imputed variables[a]	358	7,145	5.01
C_DisAge_i	Age at onset of disability	48 longitudinal imputation, 221 hot deck	269	8,824	3.05
C_AdultChildOnset_i	Adult/child onset of disability	26 constructed from imputed variables	26	8,824	0.29
B24_i	Currently working	6 hot deck	6	8,824	0.07

Source: NBS Round 7.

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

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

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

3. Section C: Current jobs variables

Several survey questions asked respondents about current employment. Section C asked such questions only of respondents who indicated in Item B24 that they were currently working. If the respondent did not answer Item B24, then we imputed Item B24. In this round, there were six such cases, four 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, 7 hot deck	8	4,364	0.18
C2_1_1d_i	Main current job SOC code to one digit	15 hot deck[a]	15	4,364	0.34
C8_1_i	Hours per week usually worked at current main job	67 hot deck,[b] 4 imputed by distributional assumptions	71	4,364	1.62
C_TotCurWkHrs_i	Total weekly hours at all current jobs	67 hot deck,[c] 14 constructed from imputed variables	81	4,364	1.86
C_TotCurHrMnth_i	Total hours per month at all current jobs	77 constructed from imputed variables	77	4,364	1.76
C_MainCurJobHrPay_i	Hourly pay at current main job	10 logical, 390 constructed from imputed variables	400	4,364	9.17
C_MainCurJobMnthPay_i	Monthly pay at current main job	36 logical, 26 imputed by distributional assumptions, 364 constructed from imputed variables	426	4,364	9.76
C_TotCurMnthPay_i	Total monthly salary all current jobs	33 logical, 364 hot deck, 44 constructed from imputed variables	441	4,364	10.11

Source: NBS Round 7.

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	24 hot deck	24	8,824	0.27
I9_i	Current health	68 hot deck	68	8,824	0.77
I17b_i	Blind or difficulty seeing, even with glasses	2 logical, 104 hot deck	106	8,824	1.20
I19_i	Uses special equipment because of difficulty seeing	12 hot deck, 89 constructed from imputed variables	101	8,824	1.14
I21_i	Deaf or difficulty hearing	1 logical, 94 hot deck	95	8,824	1.08
I22_i	Able to hear normal conversation at all	32 hot deck, 81 constructed from imputed variables	113	8,824	1.28
I23_i	Uses special equipment because of difficulty hearing	13 hot deck, 81 constructed from imputed variables	104	8,824	1.18
I25_i	Difficulty having speech understood	6 logical, 110 hot deck	116	8,824	1.31
I26_i	Able to have speech understood at all	37 hot deck, 85 constructed from imputed variables	122	8,824	1.38
I27_i	Uses special equipment because of difficulty speaking	19 hot deck, 85 constructed from imputed variables	104	8,824	1.18
I29_i	Difficulty walking or climbing stairs without assistance	3 logical, 98 hot deck	101	8,824	1.14
I30_i	Able to walk without assistance at all	65 hot deck, 48 constructed from imputed variables	113	8,824	1.28
I31_i	Uses special equipment because of difficulty walking	48 hot deck, 48 constructed from imputed variables	96	8,824	1.08
I34_i	Able to climb stairs at all	73 hot deck, 48 constructed from imputed variables	121	8,824	1.37
I35_i	Difficulty lifting and carrying 10 pounds	1 logical, 113 hot deck	114	8,824	1.29
I36_i	Able to lift or carry 10 pounds at all	85 hot deck, 73 constructed from imputed variables	158	8,824	1.79
I37_i	Difficulty using hands or fingers	116 hot deck	116	8,824	1.31
I38_i	Able to use hands or fingers at all	47 hot deck, 86 constructed from imputed variables	133	8,824	1.50
I39_i	Difficulty reaching over head	1 logical, 116 hot deck	117	8,824	1.32
I40_i	Able to reach over head at all	42 hot deck, 86 constructed from imputed variables	128	8,824	1.45
I41_i	Difficulty standing	1 logical, 127 hot deck	128	8,824	1.45
I42_i	Able to stand at all	67 hot deck, 56 constructed from imputed variables	123	8,824	1.39
I43_i	Difficulty stooping	3 logical, 111 hot deck	114	8,824	1.29
I44_i	Able to stoop at all	80 hot deck, 54 constructed from imputed variables	134	8,824	1.52
I45_i	Difficulty getting around inside home	5 logical, 111 hot deck	116	8,824	1.32
I46_i	Needs help to get around inside home	24 hot deck, 93 constructed from imputed variables	117	8,824	1.32

Table IV.4 (*continued*)

Variable name	Description	Imputation method	Number missing	Number eligible	Percentage imputed
I47_i	Difficulty doing errands alone	3 logical, 115 hot deck	118	8,824	1.33
I48_i	Needs help to get around outside home	85 hot deck, 64 constructed from imputed variables	149	8,824	1.69
I49_i	Difficulty getting into/out of bed	5 logical, 120 hot deck	125	8,824	1.42
I50_i	Needs help getting into/out of bed	35 logical, 91 hot deck, constructed from imputed variables	126	8,824	1.43
I51_i	Difficulty bathing or dressing	6 logical, 125 hot deck	131	8,824	1.49
I52_i	Needs help bathing or dressing	31 hot deck, 97 constructed from imputed variables	128	8,824	1.45
I53_i	Difficulty shopping	18 logical, 111 hot deck	129	8,824	1.46
I54_i	Needs help shopping	41 hot deck, 78 constructed from imputed variables	119	8,824	1.34
I55_i	Difficulty preparing own meals	11 logical, 122 hot deck	133	8,824	1.50
I56_i	Needs help to prepare meals	45 hot deck, 86 constructed from imputed variables	131	8,824	1.48
I57_i	Difficulty eating	1 logical, 116 hot deck	117	8,824	1.32
I58_i	Needs help to eat	17 hot deck, 99 constructed from imputed variables	116	8,824	1.31
I59_i	Trouble concentrating or remembering	148 hot deck	148	8,824	1.68
I60_i	Trouble coping with stress	179 hot deck	179	8,824	2.03
I61_i	Trouble getting along with people	167 hot deck	167	8,824	1.89
CageScore_Indicator_i	CAGE Alcohol Score	125 constructed from imputed variables	125	8,824	1.42
I72_i	Uses drugs in larger amounts than prescribed	150 hot deck	150	8,824	1.70

Source: NBS Round 7.

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	90 constructed from imputed variables	90	8,824	1.02
C_NumSenLim_i	Number of sensory limitations	142 constructed from imputed variables	142	8,824	1.61
C_NumSevSenLim_i	Number of severe sensory limitations	127 constructed from imputed variables	127	8,824	1.44
C_NumPhyLim_i	Number of physical functional limitations	207 constructed from imputed variables	207	8,824	2.35
C_NumSevPhyLim_i	Number of severe physical functional limitations	262 constructed from imputed variables	262	8,824	2.97
C_NumEmotLim_i	Number of emotional/social limitations	255 constructed from imputed variables	255	8,824	2.89
C_NumADLs_i	Number of impaired ADL	173 constructed from imputed variables	173	8,824	1.96
C_NumADLAssist_i	Number of ADL requiring assistance	145 constructed from imputed variables	145	8,824	1.64
C_NumIADLs_i	Number of IADL difficulties	171 constructed from imputed variables	171	8,824	1.94
C_NumIADLAssist_i	Number of IADL requiring assistance	171 constructed from imputed variables	171	8,824	1.94
C_PCS8TOT_i	Physical summary score	237 constructed from imputed variables	237	8,824	2.69
C_MCS8TOT_i	Mental summary score	237 constructed from imputed variables	237	8,824	2.69
C_DrugDep_i	Drug dependence	154 constructed from imputed variables	154	8,824	1.75

Source: NBS Round 7.

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	231 constructed from imputed variables, 24 imputed by descriptive statistics using specialized procedures	255	8,824	2.91
C_AmtWorkComp_i	Amount received from workers' compensation last month	154 constructed from imputed variables, 7 imputed by descriptive statistics using specialized procedures	161	8,824	1.83
C_AmtVetBen_i	Amount received from veterans' benefits last month	144 constructed from imputed variables, 20 imputed by descriptive statistics using specialized procedures	164	8,824	1.86
C_AmtPubAssis_i	Amount received from public assistance last month	151 constructed from imputed variables, 18 imputed by descriptive statistics using specialized procedures	169	8,824	1.91
C_AmtUnemply_i	Amount received from unemployment benefits last month	142 constructed from imputed variables, 3 imputed by descriptive statistics using specialized procedures	145	8,824	1.64
C_AmtPrivPen_i	Amount received from private pension last month	146 constructed from imputed variables, 15 imputed by descriptive statistics using specialized procedures	161	8,824	1.82
C_AmtOthReg_i	Amount received from other regular sources last month	151 constructed from imputed variables, 18 imputed by descriptive statistics using specialized procedures	169	8,824	1.91

Source: NBS Round 7.

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 any prior rounds or in Round 7, 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. The programming error was corrected in Round 7, so that the missingness in the C_Cohab variable in Round 7 (1.80 percent) was more in line with what had been observed in Rounds 1 to 5.

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[82] 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 was specific to the variable being imputed.

[82] 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	432 hot deck	432	8,824	4.90
L3_i	Highest year/grade completed in school	198 hot deck	198	8,824	2.24
L8_i	Marital status	179 hot deck	179	8,824	2.03
L11_i	Living arrangements	7 logical, 165 hot deck	172	8,824	1.95
C_NumChildHH_i	Number of children living in household	18 logical, 156 hot deck, 42 constructed from imputed variables	216	8,824	2.45
C_HHsize_i	Household size	1 logical, 179 hot deck, 31 constructed from imputed variables	211	8,824	2.39
C_Cohab_i	Cohabitation status	6 logical, 153 hot deck	159	8,824	1.80
C_FedPovertyLevel_cat	2018 federal poverty level	3,322 constructed from imputed variables	3,322	8,824	37.65

Source: NBS Round 7.

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 statistical procedures in 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.[83]

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[84] 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

[83] 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.hcp.med.harvard.edu/statistics/survey-soft/survey-soft. 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.

[84] 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).

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

[85] 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.

Callahan, R., E. Grau, A. Wec, K. McDonald, B. Mory, L. Pranschke, and J. Markesich. "The National Beneficiary Survey-General Waves Round 7 (Volume 3 of 3): User's Guide for Restricted and Public Use Data Files." Washington, DC: Mathematica, 2021.

Callahan, R., K. McDonald, J. Markesich and G. Livermore. "The National Beneficiary Survey-General Waves Round 7 Questionnaire." Washington, DC: Mathematica, 2021.

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., and H. Zhou. "The National Beneficiary Survey–General Waves: Round 7: Nonresponse Bias Analysis." Washington, DC: Mathematica, 2021.

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., A. Wec, R. Callahan, J. Markesich, B. Mory, and E. Grau. "The National Beneficiary Survey—General Waves Round 7: Public-Use File Codebook." Washington, DC: Mathematica, 2021.

McDonald, K., B. Mory, R. Callahan, A. Wec, and J. Markesich. "The National Beneficiary Survey—General Waves Round 7: Restricted-Use File Codebook." Washington, DC: Mathematica, 2021.

McDonald, K., R. Callahan, A. Wec, B. Mory, L. Pranschke, E. Grau, and J. Markesich. "National Beneficiary Survey—General Waves Round 6 (Volume 2 of 3): Data Cleaning and Identification of Data Problems." Washington, DC: Mathematica, 2021.

O'Day, B., Hannah Burak, K. Feeney, E. Kelley, F. Martin, G. Freeman, G. Lim, and K. Morrison. "Employment and Experiences of Young Adults and High Earners Who Receive Social Security Disability Benefits: Findings from Semi-Structured Interviews." Washington, DC: Mathematica Policy Research, March 2016.

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 9.1 User's Guide. Cary, NC: SAS Institute, 2017.

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

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.1. "Other/Specify" and Open-Ended Items with Additional Categories Used During Coding

Question #	Question Text	Current Response Options	Additional Categories Used
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

Question #	Question Text	Current Response Options	Additional Categories Used
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
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

Question #	Question Text	Current Response Options	Additional Categories Used
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)
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?

Question #	Question Text	Current Response Options	Additional Categories Used
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
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)

Question #	Question Text	Current Response Options	Additional Categories Used
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?
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

Question #	Question Text	Current Response Options	Additional Categories Used
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
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 2018? (SPECIFY: <OPEN>)	g=Had medical problems/complications

Question #	Question Text	Current Response Options	Additional Categories Used
D25_2	What benefits were reduced or ended as a result of [your/NAME's] job in 2018?	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
D26_h	In 2018, 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 2018? (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)

Question #	Question Text	Current Response Options	Additional Categories Used
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)

Question #	Question Text	Current Response Options	Additional Categories Used
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 2018.	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 2018. 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
G22	Where did {you/NAME} receive this mental health therapy or counseling? Please think about all of the places {you/NAME} went in 2018. 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

Question #	Question Text	Current Response Options	Additional Categories Used
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.1. 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 Specialists
132	Financial Specialists
Computer and Mathematical Science	
151	Computer Occupations
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
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

Code	Occupation
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
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

Code	Occupation
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
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

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

Appendix C

NAICS Industry Codes

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

Table C.1. 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**
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
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

Code	Description
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, Book, and Music 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)
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

Code	Description
55	**Management of Companies and Enterprises**
551	Management of Companies and Enterprises
56	**Administrative and Support and 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
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

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

Appendix D

Parameter Estimates and Standard Errors for Nonresponse Models

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

Table D.1. Variables in the location logistic propensity model in the RBS

Main effects	Parameter estimate[a]	Standard error
Variables in the location model, Representative Beneficiary Sample		
Number of phone numbers on file (PHONE)		
One	0.692	0.471
Two	0.477	0.494
Three	1.325**	0.502
Four	1.123*	0.527
Five or more, or zero	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.416	0.275
Age in range 30 to 39 years	-0.308	0.240
Age in range 40 to FRA	Ref. cell	
U.S. Census division (DIVISION)		
Middle Atlantic	-0.619*	0.265
Not Middle Atlantic	Ref. cell	
Beneficiary's race (RACE)		
White	-1.050**	0.320
Black	-0.511	0.375
Not White or Black	Ref. cell	
Identify of payee relative to beneficiary (REPREPAYEE)		
Family	0.526	0.353
Not family	Ref. cell	
Beneficiary title (SSI_SSDI) (ONLY SSI)		
SSI only	-0.720*	0.358
Other	Ref. cell	
Retirement destination county (CNTYRET)		
The number of residents in county age 60 and older grew by 15 percent or more between the 2000 and 2010 censuses due to net migration	0.736*	0.316
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 RBS

Main Effects	Parameter estimate[a]	Standard error
Variables in the cooperation model, Representative Beneficiary Sample		
Number of addresses on file (MOVE)		
One	0.490*	0.206
Two	0.333	0.223
Three	0.153	0.215
Four	0.427	0.258
Five or more, or zero	Ref. cell	
Ethnicity (HISPANIC)		
Hispanic	0.601*	0.230
Not Hispanic	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.106	0.126
Age in range 30 to 39 years	-0.134	0.127
Age in range 40 to 49 years	-0.037	0.128
Age in range 50 to FRA	Ref. cell	
Gender (GENDER)		
Female	0.158	0.115
Male	Ref. cell	
County with high levels of children living in poverty (CNTYCPOV)		
Yes	0.806**	0.262
No	Ref. cell	
County with high levels of persistent poverty (CNTYPPOV)		
Yes	-0.546	0.282
No	Ref. cell	
County with recreation-based economy (CNTYREC)		
Yes	0.282	0.169
No	Ref. cell	
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in nonmetropolitan area	0.511**	0.143
Beneficiary resides in metropolitan statistical area (MSA) of less than 250,000	0.235	0.193
Beneficiary resides in metropolitan statistical area (MSA) of 250,000 or more	Ref. cell	
Earnings category (EARNCAT)		
Monthly DCF earnings above SGA[c] for three consecutive months in 2017 or 2018	-0.489	0.313
Gross annual DCF earnings above three times SGA in 2017 or 2018	0.304	0.294
Gross annual DCF earnings above $0 in 2017 or 2018	0.218	0.242
No annual DCF earnings in 2017 or 2018	Ref. cell	
Two-Factor Interactions[b]		

Table D.2 (*continued*)

Main Effects	Parameter estimate[a]	Standard error
CNTYCPOV * AGECAT		
County with high levels of child in poverty * Age in range 18 to 29	-0.846**	0.250
County with high levels of children in poverty * Age in range 30 to 39	-0.206	0.265
County with high levels of children in poverty * Age in range 40 to 49	-1.016**	0.282
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

Table D.3. Variables in the location logistic propensity model in the cross-sectional SWS

Main effects	Parameter estimate[a]	Standard error
Variables in the location model, Successful Worker Sample		
Extract (EXTRACT)		
First extract	0.612**	0.193
Second extract	0.351	0.197
Third extract	-0.235	0.206
Fourth extract	0.000	0.182
Fifth extract	-0.362*	0.174
Sixth extract	-0.226	0.181
Seventh extract	Ref. cell	
Number of phone numbers on file (PHONE)		
One	-0.663**	0.164
Two	-0.348*	0.156
Three	0.097	0.151
Four	-0.040	0.148
Five or more, or zero	Ref. cell	
Number of addresses on file (MOVE)		
One	-0.276	0.150
Two	-0.021	0.131
Three or more, or zero	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.282*	0.116
Age in range 30 to FRA	Ref. cell	
Beneficiary's living situation (LIVING)		
Beneficiary lives alone	0.342	0.245
Beneficiary lives with family, others, in an institution, or situation unknown	Ref. cell	
County with government-dependent economy (CNTYGOV)		
Yes	0.431**	0.163
No	Ref. cell	
Beneficiary title (SSI_SSDI)		
SSDI only recipient	1.027**	0.366
Recipient of SSI (concurrent or SSI only)	Ref. cell	
County with nonspecialized-dependent economy (CNTYNONSP)		
County with nonspecialized-dependent economy	0.312**	0.119
County that doesn't have this attribute	Ref. cell	
Earnings category (EARNCAT)		
Monthly DCF earnings above SGA[b] for three consecutive months in 2017 or 2018	-0.069	0.217
Gross annual DCF earnings above three times SGA in 2017 or 2018	0.200	0.281
Gross annual DCF earnings above $0 in 2017 or 2018	-0.414	0.319
No annual DCF earnings in 2017 or 2018	Ref. cell	

Table D.3 (*continued*)

Main effects	Parameter estimate[a]	Standard error
Indicator whether beneficiary and applicant for benefits are in same zip code (PDZIPSAME)		
Applicant and beneficiary live in same zip code	1.188**	0.347
Applicant and beneficiary live in different zip code	1.134**	0.373
Unknown	Ref. cell	
Beneficiary's race		
Non-Hispanic Black	0.486**	0.120
Not non-Hispanic black, or race unknown	Ref. cell	
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in metropolitan statistical area (MSA) of less than 250,000	-0.232	0.134
Beneficiary resides in nonmetropolitan area adjacent to large metropolitan area	0.449	0.261
Beneficiary resides in nonmetropolitan area adjacent to small, medium, or no metropolitan area	0.416	0.251
Beneficiary resides in metropolitan statistical area (MSA) of more than 250,000	Ref. cell	
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in nonmetropolitan area not adjacent to metropolitan area	-0.232	0.134
Beneficiary resides in nonmetropolitan area adjacent to large metropolitan area	0.449	0.261
Beneficiary resides in nonmetropolitan area adjacent to medium or small metropolitan area, or not adjacent to a metropolitan area	0.416	0.251
Beneficiary resides in metropolitan statistical area (MSA) of more than 250,000	Ref. cell	
Two-factor interactions[b]		
LIVING * MOVE		
Not living alone * One address	-0.510*	0.229
Not living alone * Two addresses	0.424	0.228
Successful worker missing one or both of these attributes	Ref. cell	
RACE * MOVE		
Not non-Hispanic black * One address	0.837**	0.240
Not non-Hispanic black * Two addresses	0.515	0.264
Successful worker 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

Table D.4. Variables in the cooperation logistic propensity model in the cross-sectional SWS

Main Effects	Parameter estimate[a]	Standard error
Variables in the cooperation model, Successful Worker Sample		
Extract (EXTRACT)		
First extract	0.790**	0.194
Second extract	0.483**	0.130
Third extract	0.151	0.131
Fourth extract	0.111	0.136
Fifth extract	-0.056	0.147
Sixth extract	0.135	0.129
Seventh extract	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.231	0.168
Age in range 30 to 39 years	-0.368**	0.074
Age in range 40 to 49 years	-0.225*	0.091
Age in range 50 to FRA	Ref. cell	
Beneficiary's disability category (DISABILITY)		
Deafness	-0.475*	0.219
Other disability excluding deafness, or disability unknown	Ref. cell	
Identity of payee relative to beneficiary (REPREPAYEE)		
Beneficiary received payments himself/herself	0.222*	0.111
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	0.202**	0.066
Applicant and beneficiary live in different zip code, or no information	Ref. cell	
DCF earnings category in 2017-2018 (EARNCAT)		
Monthly DCF earnings above SGA[b] for three consecutive months in 2017 or 2018	0.029	0.138
Gross annual DCF earnings above three times SGA in 2017 or 2018	0.177	0.175
Gross annual DCF earnings above $0 in 2017 or 2018	0.463*	0.226
No annual DCF earnings in 2017 or 2018	Ref. cell	
ETHNICITY		
Hispanic	-0.375**	0.140
Not Hispanic	Ref. cell	
County with high levels of poverty (CNTYHPOV)		
County with high levels of poverty	0.267	0.136
County that doesn't have this attribute	Ref. cell	
Two-Factor Interactions[b]		
AGECAT * EXTRACT		
Age in range 30 to FRA not in EXTRACT1	0.386*	0.188
Beneficiary missing one or both of these attributes	Ref. cell	

Table D.4 (*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.5. Variables in the location logistic propensity model in the longitudinal SWS, in Round 7 beneficiary frame

Main effects	Parameter estimate[a]	Standard error
Variables in the location model, Successful Worker Sample		
Extract (EXTRACT)		
Fifth extract	0.381	0.220
Sixth extract	0.538	0.282
First through fourth or seventh extract	Ref. cell	
Number of phone numbers on file (PHONE)		
Zero	1.056**	0.384
One	0.373	0.164
Two	0.163	0.156
Three	-0.192	0.151
Four	-0.008	0.148
Five or more	Ref. cell	
U.S. Census region (REGION)		
Midwest	-0.428*	0.186
West	-0.717**	0.174
South	-0.479*	0.227
Northeast	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.952	0.095
Age in range 30 to 39 years	-1.079*	0.014
Age in range 40 to 49 years	-0.040	0.945
Age in range 50 to FRA	Ref. cell	
Beneficiary's race (RACE)		
Black	0.409*	0.193
Not black, 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	-0.951**	0.344
Applicant and beneficiary live in different zip code, or no information	Ref. cell	
Beneficiary title (SSI_SSDI)		
SSDI only recipient	1.068**	0.349
Recipient of SSI (concurrent or SSI only)	Ref. cell	
Metropolitan status of county of residence of beneficiary (METRO)		
Beneficiary resides in metropolitan statistical area (MSA) of less than 250,000	-0.923**	0.250
Beneficiary resides in metropolitan statistical area (MSA) of 250,000-999,999	-0.261	0.183
Beneficiary resides in metropolitan statistical area (MSA) of 1 million or more, or in nonmetropolitan area	Ref. cell	
County with government-dependent economy (CNTYGOV)		
Yes	0.295	0.419
No	Ref. cell	

Table D.5 (*continued*)

Main effects	Parameter estimate[a]	Standard error
Categorized percentage of housing units in county that do not use fuel (CNTYNOFUEL)		
Less than 0.4 percent	0.286	0.178
Between 0.4 and 0.6 percent	0.403	0.210
More than 0.6 percent	Ref. cell	
Two-factor interactions[b]		
CNTYGOV * AGECAT		
Not a government-dependent economy * Age 18 to 29	0.876	0.611
Not a government-dependent economy * Age 30 to 39	0.999*	0.503
Not a government-dependent economy * Age 40 to 49	-0.125	0.632
Successful worker 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

Table D.6. Variables in the cooperation logistic propensity model in the longitudinal SWS, in Round 7 frame

Main Effects	Parameter estimate[a]	Standard error
Variables in the cooperation model, Successful Worker Sample		
Extract (EXTRACT)		
First extract	0.270*	0.133
Third extract	0.673*	0.297
Seventh extract	0.438*	0.170
Second, fourth, fifth, or sixth extract	Ref. cell	
Number of addresses on file (MOVE)		
One	0.544	0.342
Two	-0.131	0.357
Three	0.177	0.308
Four or more	Ref. cell	
Beneficiary's age category (AGECAT)		
Age in range 18 to 29 years	-0.634**	0.129
Age in range 30 to 39 years	-0.584**	0.121
Age in range 40 to 49 years	-0.242*	0.122
Age in range 50 to FRA	Ref. cell	
Beneficiary's race (RACE)		
Non-Hispanic Black	0.206	0.128
Not non-Hispanic black, or race unknown	Ref. cell	
Beneficiary's living situation (LIVING)		
Beneficiary lives with others	-0.503	0.295
Beneficiary lives with family, others, in an institution, or situation unknown	Ref. cell	
Beneficiary title (SSI_SSDI)		
Recipient of SSDI and SSI	0.820*	0.360
Recipient of SSI (concurrent or SSI only)	Ref. cell	
U.S. Census region or division (REGION or DIVISION)		
Middle Atlantic	0.368**	0.120
West	0.280*	0.126
South	0.235*	0.118
Northeast	Ref. cell	
Retirement destination county (CNTYRET)		
The number of residents in county age 60 and older grew by 15 percent or more between the 2000 and 2010 censuses due to net migration	-0.430**	0.166
County does not have this attribute	Ref. cell	
County that doesn't have this attribute		
Two-factor interactions[b]		
SSI_SSDI * EXTRACT		
Extracts 1, 2, 4-7 * SSI only or SSDI only	-0.484	0.320
Beneficiary missing one or both of these attributes	Ref. cell	

Table D.6 (*continued*)

Main Effects	Parameter estimate[a]	Standard error
SSI_SSDI * MOVE		
One address * SSI only or SSDI only	-0.551	0.365
Two addresses * SSI only or SSDI only	-0.107	0.370
Three or more addresses * SSI only or SSDI only	-0.559	0.322
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: Wtr7_ben
Cross-sectional SWS: Wtr7_cssws
Longitudinal SWS: Wtr7_lngsws
Combined samples: Wtr7_com

NEST VARIABLES USED FOR CROSS-SECTIONAL ESTIMATES

A_STRATA

1. Clustered samples for RBS and cross-sectional 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.

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.org