2015-2017 National Survey of Family Growth (NSFG): Sample Design Documentation

This document is a detailed supplement to another document that serves as a brief, summary description of all aspects of the methodology and operations for the 2015-2017 data release. The summary document is referred to as "summary methodology document" below and is entitled "2015-2017 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods".

1. Executive Summary

The National Survey of Family Growth (NSFG) is a survey on fertility, family formation and change, family planning, reproductive health, and closely related topics. It is a principal source of national estimates on a variety of fertility and family topics. The target population for the NSFG consists of all non-institutionalized women and men ages 15-49 years as of first contact for the survey, living in households, and whose usual place of residence is the 50 United States and the District of Columbia.

Analogous to the 2006-2010 NSFG (Lepkowski et al., 2013), the sample design for the 2011-2019 NSFG survey period was developed to be a continuous design with independent two-year samples being released periodically. These nationally representative samples can be combined for differing time intervals. This document describes the sampling procedures for the development of the sample for the time period 2015-2017. This two-year period covers the 17th through 24th quarters of the overall planned eight years of data collection for 2011-2019. See "2011-2013 National Survey of Family Growth (NSFG): Sample Design Documentation" and "2013-2015 National Survey of Family Growth (NSFG): Sample Design Documentation" for analogous reports for the first eight quarters (September 2011 through September 2013) and the second eight quarters (September 2013 through September 2015) comprising the first and second data releases from continuous interviewing. One significant change was made for the

2015-2017 sample design. Beginning in September 2015, or quarter 17 of fieldwork, the survey's age eligibility was expanded from 15-44 years to 15-49 years.

The sample design for the NSFG 2011-2019 survey period is based on goals of completing a minimum of 5,000 interviews per year with significant oversamples of non-Hispanic blacks, Hispanics, and teens aged 15-19, and a slightly higher sampling rate for females versus males. Specifically, the objectives call for 20% of the interviews to be with non-Hispanic blacks, 20% with Hispanics, 20% with teens (15-19 years of age), and 55% with females. The goal for teens was reduced to 18.2% in order to accommodate the inclusion of additional adults (i.e., those 45-49 years of age). This goal was set in order to ensure a sufficient sample of teens, given that most households containing a teen also contained an adult(s): the algorithm selected teens at a higher rate than adults when both were present in the household. Regarding Hispanics and non-Hispanic blacks, the percentage of households in the population containing those racial ethnic groups around that time period were 17.5% and 14.7%, respectively. The oversample objectives were achieved through a number of measures. These measures, and the changes to the rates of selection for teens, are described below.

The sample was selected in five stages. In the **first stage**, Primary Sampling Units (PSUs) were selected. PSUs are Metropolitan Statistical Areas (MSAs), counties or groups of counties. The 50 United States plus the District of Columbia were divided into 2,149 PSUs on the sampling frame. Of these, 366 were MSAs and 1,783 were non-MSA PSUs that include one or more counties. The PSUs were stratified according to attributes such as Census Division, MSA status, and size. One or two PSUs were selected with Probability Proportionate to Size (PPS) from each stratum. The PPS selection method assigns higher probabilities to PSUs with larger populations. The first stage selection probabilities are inversely related to the probabilities of selection at the second and third stages of selection such that sampling rates are approximately equal for all households within a sampling domain (defined below). In general, large PSUs have lower within-PSU sampling rates while smaller PSUs have higher within-PSU sampling rates such that households in the same domain but different PSUs have approximately the same chance of being selected. The largest PSUs were selected with probability equal to 1.0 since any national sample of this size should include them. These PSUs are known as "certainty selections" or "self-representing" PSUs. These self-representing PSUs are in strata with only one PSU per stratum. There are 21 such self-representing strata. The remaining PSUs were grouped into strata of approximately the same size. There are an additional 192 non-self-representing strata plus two strata for Alaska and Hawaii.

In order to facilitate the oversample of subgroups defined by race and ethnicity, the measures of size for the PSUs were a weighted combination of household counts. All Census Block Groups were classified into four sampling "domains." Table 1 shows the definition of the four domains. Households in domains 2, 3, and 4 were given a higher weight so that they would have a higher chance of being selected than those in domain 1. These weighted measures of size were then used in both the first and second stages of selection.

Table 1. Domain Definitions and Characteristics

Domain	Definition	Total Households	Est. Proportion Black	Est. Proportion Hispanic
1	<10% HH Black, <10% HH Hispanic	65,009,685	0.018	0.022
2	>=10% HH Black, <10% HH Hispanic	19,871,976	0.426	0.029
3	<10% HH Black, >=10% HH Hispanic	20,270,438	0.026	0.380
4	>=10% HH Black, >=10% HH Hispanic	11,564,193	0.301	0.299

The **second stage** of selection was to select neighborhoods within PSUs. These selections are called Secondary Sampling Units (SSUs or segments) and are composed of one or more Census Blocks with a minimum measure of size equal to 50. The minimum size requirement ensures that within-SSU samples are large enough to support efficient travel by interviewers. SSUs were selected with PPS. The measures of size for these PPS selections are weighted measures of size such that SSUs with larger non-Hispanic black and Hispanic populations receive higher probabilities of selection.

SSUs in domains 2, 3, and 4 have relatively higher combined PSU, SSU, and housing unit selection rates. These weighted measures of size and sampling rates were set such that interviews with black and Hispanic respondents each constitute 20% of all interviews.

Each PSU was assigned one or two interviewers based on its relative size. For each interviewer, 12 SSUs were selected each year. These SSUs were then randomly divided into four groups. One group was released each quarter in a data collection year.

In preparation for the **third stage** of selection, interviewers from the Institute for Social Research (ISR) updated commercially-available lists of housing units for SSUs where these lists are available or, alternatively, created such a list from scratch where they are not available. Once these lists were updated, a sample of housing units was selected.

The selected units were contacted by ISR interviewers to determine if any members of the household were eligible. In households with eligible persons, a **fourth stage** of selection involved selecting one of the eligible persons. Prior to September 2015, the within-household selection rates were set up such that 20% of all interviews were with teens aged 15-19 and 55% of all interviews were with females. Beginning with 2015-2017 data collection, the age range was expanded to include persons 15-49 years of age. This required modification of the within-household selection probabilities and a reduction of the expected proportion of interviews with teens from 20% to 18.2%.

As was done in NSFG 2006-2010, 2011-2013 and 2013-2015, a two-phase sampling approach was used as a **fifth stage** of selection. Each quarter, during week 10, a subsample of active cases was selected for continued follow-up. In weeks 11 and 12, this subsample received a special mailed incentive and the interviewers focused effort on the fewer cases left in the subsample.

The NSFG 2015-2017 sample is a random subsample of the full 2011-2019 NSFG sample. Smaller PSUs were randomly selected for each year. Other PSUs were large enough to occur in more than one year. In this way, each year's sample is a nationally representative sample. Multiple years can be combined, but for a minimum of two-year increments. Single-year case weights were not estimated since sample sizes would not be sufficient to provide statistically reliable estimates. The 2015-2017 NSFG data release is the third release for the NSFG 2011-2019 survey period and includes the sample selected for the third two-year interval of this period.

As with NSFG 2006-2010, 2011-2013 and 2013-2015, responsive design options were available at several levels of the sample design, which were implemented as needed:

- reducing the number of PSUs at the beginning of each data collection year (i.e., the point at which PSUs rotate in and out of the sample
- changing the number of SSUs at the beginning of each quarter
- changing sampling rates for housing units at the beginning of each quarter (in response to changed estimates of interviewer efficiency)
- changing the oversampling rates for non-Hispanic blacks and Hispanics by changing the weighted measure of size used to select the SSUs
- altering the second phase sampling procedures during any quarter

2. Introduction

This document describes the methods and procedures used for the selection of a nationally representative sample of the US household population ages 15-49, with oversamples of teens, blacks, and Hispanics. The purpose of this sample selection is to serve as the basis of the NSFG fieldwork for the time period from September 2015 to September 2017. The document follows the order of selection and proceeds from stratification and selection of Primary Sampling Units, to the selection of Secondary Sampling Units, housing units, and persons within households.

2.1 Design Specifications

The sample design described in this document is based on goals for the overall 2011-2019 data collection period and included completing a minimum of 5,000 interviews per year. Specifically, the objectives called for 20% of the interviews to be with non-Hispanic blacks, 20% with Hispanics, 20% with teens (15-19 years of age), and 55% with females. Beginning in 2015, the objective for teens was lowered to 18.2% of interviews to accommodate the addition of those aged 45-49. The percentage of households in the U.S. population over the fieldwork time

period with a household member who was Hispanic was 17.5, and percentage of households containing a member who was non-Hispanic black was 14.7%.

In order to accommodate hypothetical future changes in funding, wherever possible, procedures were built in for reducing the sample design. Since each year is an independent national sample, the number of years can be reduced. In addition, the within-PSU sampling rates can be changed each year. Details regarding the scalability of the design are discussed in Section 8, "Sample Selection in a Responsive Design Framework."

2.2 =-The Sample Universe

The survey population for the NSFG 2015-2017 consists of all non-institutionalized women and men ages 15-49 years, whose usual place of residence is the 50 United States and the District of Columbia. Excluded from the survey population are those in institutions, such as prisons, homes for juvenile delinquents, homes for the intellectually disabled, and long-term psychiatric hospitals, and those living on military bases. Age eligible persons living in noninstitutional group quarters (e.g., dormitories, fraternities) are specifically included; college students living in dormitories, fraternities, or sororities were sampled through their parent or guardians' households. In addition, women and men who are in the military but living off base are part of the NSFG sample.

2.3 The Sample Design

The sample design for the NSFG 2015-2017 is a stratified multi-stage area probability sample. In a series of steps, geographically defined sampling units of decreasing size were selected with probability proportionate to size. This design has five stages of selection: a first-stage selection of Metropolitan Statistical Areas, counties and county groups; a second-stage selection of neighborhoods defined by Census Blocks; a third-stage selection of housing units; and a fourth-stage selection of persons within households. A second-phase sample was drawn during the field period to address nonresponse. This document follows the order in which the selections were made.

The aforementioned oversampling of gender, age, and race/ethnicity groups was accomplished by screening a sample of households. Oversampling of race and ethnic subgroups was done as part of the first and second stages of selection. The oversampling of teens and females was done as part of the within-household selection procedure. The design explicitly controlled for completed interview targets by individual gender, age, race, and ethnicity groups – not simply by the marginal distributions of the cross-classification of interview outcomes.

2.4 NSFG Sampling Frame

Full coverage of the eligible subpopulations of the NSFG survey population is essential to minimize the total survey error of resulting statistics. The NSFG 2015-2017 sample used a stratified multi-stage area sample frame for U.S. households. This frame combines comprehensive data from the US Census Bureau for the Metropolitan Statistical Area (MSA), county, tract, block group, and block levels with updated TIGER (**T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing system) databases defining the boundaries of these units.

The frame also included housing unit lists within selected segments, with segments defined as either a single block or a group of blocks selected together to provide a minimum of at least 50 occupied housing units, obtained either from a "scratch" field listing or from a commercial vendor of the U.S. Postal Service Delivery Sequence File (DSF) addresses for the selected blocks. The latter DSF addresses were checked manually in the field to delete incorrectly assigned units and add missed housing units.

The last stage of selection, within-household sampling, was completed by trained ISR interviewing staff who attempted to complete a listing of all members of each sampled household. A computerized selection routine then selected one eligible person per household to be interviewed.

The NSFG 2015-2017 Primary Stage Units (PSUs, described below) were selected using the 2005-2009 American Community Survey Summary File, which was available in mid-December 2010. The Secondary Stage Units (SSUs, described below) were selected using the Census 2010 Redistricting Data [P.L. 94-171], which were available as of April 1st, 2011, and the associated geographic boundary files (also called "shapefiles").

2.5 Interviewer Labor Model

The labor model used for the 2015-2017 NSFG is consistent with the model that was used in the first four years of this data collection period (2011-2015) and that was first implemented in the 2006-2010 NSFG, upon the switch from periodic to continuous interviewing. The following is a brief description of the background and rationales.

In many U.S. survey organizations, interviewers are part-time employees working 15-20 hours per week. If an organization has periodic rather than continuous surveys, interviewers work until the end of study data collection, and then move to another organization, or wait for the next survey from the same organization. This was the labor model for ISR's implementation of NSFG Cycle 6 (2002), where interviewers were employed for at most 11 months, and there was significant staff attrition over the 11 months of data collection. The attrition, substantial training cost, the extended data collection period for a sample to be managed, and the difficulty of recruiting, hiring, and training a more highly educated interviewer work force were among the reasons for the switch from periodic to continuous interviewing in the NSFG.

Both a new labor model and a new data collection period were implemented along with continuous interviewing for the 2006-2010 and 2011-2015 NSFG. Interviewers were employed for an expected 30 hours per week continuously throughout a one-year period and data collection on a sample shortened from 11 to 3 months. Emphasis was placed on reducing the number of individual staff that needed to be managed, completing screener interviews as early in the three-month period as possible, increasing the share of interviewer time devoted to interviewing (as opposed to administrative tasks), providing a varied work assignment (for example, having interviewers update or "scratch" list second-stage units for the next quarter during data collection on a current quarter), and a second-phase sample for nonresponse.

This labor model led to lower attrition and increased efficiency during the 2006-2010, 2011-2013 and 2013-2015 NSFGs, compared to the 2002 NSFG, and yielded comparable data quality relative to the 2002 NSFG (for example, response rates were slightly lower, but target interview goals were met). This labor model was also used during the 2015-2017 NSFG and informed decisions about the second and third stages of selection.

3. Primary Stage: Design and Selection

This section describes the formation of strata and primary stage units (PSUs), first by describing the development of "weighted measures of size" and the definition of PSUs as the basis of this stage of selection. The section concludes with a general description of the stratification, PSU selections, and subsampling procedures to define a sample for the time period 2015-2017. This discussion is necessarily very general as the names of specific PSUs cannot be disclosed, due to concerns for confidentiality of survey respondents.

3.1 Weighted Measure of Size

The PSUs of this multi-stage area probability sample were selected with Probabilities Proportionate to Size (PPS). A weighted Measure of Size (MOS) is a measure where subpopulations for which an oversample is desired are multiplied by a weighting factor that increases the probability of selection for units in that domain. This allows us to oversample particular subgroups in the population. A weighted measure of size $M_{h\alpha\beta}$ for the β^{th} block in the α^{th} PSU in stratum *h* (defined below) was created as follows. If a block is in a block group where at least some threshold proportion of the population is black or Hispanic, the count of occupied housing units in that block is multiplied by a factor set such that targeted oversamples for blacks and Hispanics would be achieved. Four domains were defined (see Table 2). For blocks in Domain 1, the measure of size $M_{h\alpha\beta}$ is the 2010 Census occupied housing unit count for the block. For blocks in the other domains, the measure of size is the 2010 Census occupied housing unit count multiplied by the factors listed in the last column of Table 2.

In order to implement this weighted measure of size, cutoff values to define the domains and, subsequently, the relative sampling rates within each domain required to achieve the targeted sample sizes, needed to be determined. In order to determine appropriate cutoff values for defining the domains, an analysis of data from the American Community Survey (ACS) 2005-2009 was conducted to determine optimal boundaries for the domains. It was found that the cutoff values used for the 2002 and 2006-2010 NSFGs were still appropriate. The definition of the domains and their characteristics are presented in Table 1 above.

The weighted MOS is meant to allow oversampling of PSUs and, more importantly, SSUs with higher proportions of black and Hispanic households. Block groups in domains 2, 3, and 4 received a higher weight when calculating the measure of size for each PSU and SSU. This increased the probability of selection for areas with higher proportions of black and Hispanic households.

In many sample designs the choice of weight for each domain would simply reflect the desired relative sampling rates across the domains. In the NSFG continuous design, however, sampling rates may be raised or lowered within each PSU based on the actual productivity demonstrated by each interviewer. Therefore, the variation of sampling rates was examined across the four domains used for 2006-2010 NSFG. Table 2 presents the median relative sampling rates used in NSFG 2006-2010 by domain as well as the relative rates planned for NSFG 2011-2019.

Table 2. 2006-2010 NSFG Median Rates and Relative Sampling Rates and Planned 2011-2019 Relative Rates by Domain

Domain	2006-2010 Median Rate	2006-2010 Median Rate/Domain 1 Median Rate	2011-2019 Rate/Domain 1 Rate
1	0.000465968	1.00	1.0
2	0.001194020	2.56	2.6
3	0.000991595	2.13	2.3
4	0.001077916	2.31	2.5

The ACS 2005-2009 data were used to check expected yield from these implied sampling rates. This check included the use of race/ethnicity specific age-eligibility rates. These rates produced a sample that was somewhat less than 20% Hispanic and somewhat less than 20% black. However, selected persons from both of these groups complete the main interview, after being identified by the screening interview as eligible, at higher rates than other groups. In addition, with these groups growing as a proportion of the total population, it seemed prudent to not set oversampling rates too high. A source of information that was useful for planning the 2011-2019 sampling rates was the 2006-2010 rates. These sampling rates were set under a design that had the same targets for oversampling of black and Hispanic respondents and similar expected response rates. The relative ratio of the 2006-2010 median sampling rates to the median domain 1 rate are also presented, along with the planned relative rates set for 2011-

2019. At the end of data collection for 2015-2017, these planned rates resulted in approximately 23% of the interviews with non-Hispanic blacks, and 20% with Hispanics.

Having determined these block-level composite measures of size, the next step was to sum them to the PSU level across all blocks in the PSU to obtain a PSU level measure of size $M_{h\alpha}$, and the PSU measures of size summed to a stratum size M_h. Within a PSU stratum, a single PSU was selected with probability proportionate to the composite measure of size, or $M_{h\alpha}/M_h$. In self-representing strata, where the PSU is so large that it comes into the sample with certainty, this probability of selection is 1.0. In all other strata, it is less than 1.0.

3.2 PSU Definition

The U.S. Census Bureau provided data for all 3,143 U.S. counties and county equivalents (Louisiana parishes, Alaska boroughs and census areas, independent cities in Maryland, Virginia, Missouri, and Nevada, and the District of Columbia). The counties cover the entire land area of the U.S., and thus serve as the first level of an area frame for the U.S. household population. ISR grouped these counties into 2,149 Primary Sampling Units (PSUs), and these served as the sampling frame for the first stage selection in the continuous NSFG 2011-2019.

Counties were combined to form PSUs for two main reasons. The first was to take advantage of the Metropolitan Statistical Areas (MSAs) designated by the U.S. Office of Management and Budget (OMB). MSAs are urbanized areas that may include several counties and have a total population size of at least 50,000. MSAs are areas that have a high degree of social and economic integration. Approximately 1,100 counties are currently grouped into 366 MSAs consisting of one or more geographically adjacent counties with an urban core population of at least 50,000. The ISR PSU sample designates these MSAs as separate metropolitan sampling units. The remaining 2,043 non-metropolitan counties were treated as individual sampling units in the ISR sample. In addition to the use of multi-county MSAs, a second reason that counties were combined to form PSUs was to create areas with enough units to support the sample sizes required by the NSFG. If a county had fewer than 1,800 occupied housing units, it was linked to a neighboring county to form a single PSU with at least 1,800 housing units. This combination of MSAs, non-MSA counties, and linked groups of non-MSA counties yielded a set of 2,149 PSUs in the ISR sample frame.

The same procedures for creating PSUs were used for the 2002 NSFG and the 2006-2010 NSFG. The frame of PSUs for NSFG in 2011-2019 does not completely match the frame used for 2002 and 2006-2010. The count by MSA and non-MSA is slightly different since the definitions of MSAs change over time and the number of “small” counties that are combined to form units with a minimum measure of size has changed. However, many of the PSUs do have the same definitions across the three sample designs. PSU overlap over some of the past NSFG cycles allows some gains in precision when comparing results across cycles.

3.3 Stratification and Selection of PSUs

The general primary stage stratification is based on: a) Census Region, and where possible, Census Division, b) PSU Size, and c) PSU MSA/non-MSA status. The Census Bureau divides the U.S. into four Regions and nine Divisions. The general stratification framework can be represented by a cross-tabulation of U.S. Census Divisions and the three types of PSU categories: Self-representing (SR) MSA PSUs, Nonself-representing (NSR) MSA PSUs and non-MSA PSUs. Table 3 shows the allocation of all 213 PSUs (215 with the inclusion of PSUs selected from Alaska and Hawaii) selected for NSFG 2011-2019. Once a decision was made on the number of SR PSUs, the remaining number of strata desired for the design could be allocated to the non-SR cells in this table on the basis of total number of occupied housing units in each cell.

Table 3. Allocation of PSUs by SR vs Non-SR, Region, Division, and MSA status for NSFG 2011-2019

Region/Division	Self-Rep	Non Self-Rep Strata Counts		TOTAL
	MSA	MSA	Non-MSA	
TOTAL	21	194		215
TOTAL (omit AK & HI)	21	144	48	213
Northeast	***3***	***24***	***5***	***32***
New England	1	8	2	11
Mid Atlantic	2	16	3	21
Midwest	***4***	***26***	***11***	***41***
E No Central	2	19	6	26
W No Central	2	7	5	14
South	***7***	***62***	***26***	***95***
So Atlantic	5	32	12	49
E So Central	0	11	7	18
W So Central	2	19	7	28
West	***7***	***40***		***47***
West (omit AK & HI)	7	32	6	45
Mountain	1	13	5	19
Pacific	6	19	1	26
Alaska	0	1		1
Hawaii	0	1		1

Within the nonself-representing strata, PSU size was also a major consideration in the formation of strata; an attempt was made to group similar size PSUs within a single stratum.

The nonself-representing strata are separated by MSA/non-MSA status. Nonself-representing strata were intended to be approximately the size of the smallest self-representing stratum. A goal of the stratification was to keep each stratum within a Census division. This goal was largely met.

In eight cases, it was awkward to create a stratum from large MSAs, which would need to be combined with much smaller MSAs in order to create strata of size M_h approximately equal to 1.6 million. In these cases, “double strata” were formed, that is, strata with size of approximately 3.2 million, and then two selections were made from each of these strata. Alaska and Hawaii were treated as separate strata.

U. S. Census Regions have been maintained in this sample design as four distinct geographic domains. MSA definitions that cross these regional boundaries present a special problem. As a fictional example, an MSA that straddled the border between Kansas (in the Midwest Region) and Oklahoma (in the South Region) would cross Census Region boundaries. In order to maintain stratification by region, a total of eighteen such MSAs had to be split between PSUs in two different Census regions.

In general, this sample split these boundary-crossing MSAs to form two MSA PSUs – one PSU from the pair assigned to each of the corresponding Census regions. In the fictional example, the Kansas portion of the MSA was assigned to a stratum within the Census Midwest Region; the Oklahoma portion of the MSA was assigned to a stratum in the Census South Region. In a few cases, when the region boundary divides a primarily "rural" county from the MSA of which it is a part (due to commuting and employment patterns of its residents), it was thought that it should not stand alone as a separate MSA within its region; such separated MSA counties were, instead, linked to another nearby MSA within their Census region.

If the smaller part of a region boundary-crossing MSA had fewer than 25,000 households or if it had more than 25,000 households but no central city (by December 2009 OMB definition), then it was considered an MSA "splinter" and was linked to another MSA in its Census region and stratum. If the MSA part had more than 25,000 households and a central city, it was considered a "divided" MSA part and was allowed to stand on its own as a separate MSA in its Census region.

A sample of 213 PSUs was selected for NSFG 2011-2019 plus two PSUs to represent Alaska and Hawaii. A systematic random sample of these PSUs was drawn for each of the years of data collection. Some of the PSUs were large enough that they would be included in more than one of the years. After two years, there are nine self-representing PSUs (three included every year and six included every other year). There are a total of 12 PSUs in the 2011-2019 sample that are self-representing every three years, but have only a 2/3 chance of being selected for any two-year interval. This yields an additional eight self-representing PSUs for NSFG 2015-2017, for a total of 17 self-representing PSUs. There are 48 non-self-representing PSUs. This is a total of 65 PSUs for NSFG 2015-2017.

4. Secondary Stage: Design and Selection

The labor model determined important features of the sample design over time, including the level of sample selection within a PSU. Sufficient sample was needed each quarter to make efficient use of the time each interviewer was available. Sample size within each second-stage sampling unit was to be determined by the anticipated travel costs among second-stage units within PSUs and how many completed interviews could be expected for a given interviewer.

The 2006-2010 NSFG had approximately 12 second-stage sampling units within sample PSUs. There were more second-stage units in larger self-representing PSUs, but most PSUs had 12 second-stage units selected. These second-stage units had been selected with a minimum size of 50 housing units in urban locations and 75 in rural locations. They yielded about 16 completed interviews each from about 40 selected housing units in the unit. These sizes in turn led to design effects that were within target limits.

In the 2006-2010 NSFG, the labor model dictated the selection of a sample of second stage units within a PSU that could be conveniently divided across four quarters. It would have been possible to use fewer segments in a PSU, such as eight (two per quarter) or four (one per quarter), increasing the number of selected housing units in each to 60 or 120, respectively. However, this would have increased the within-PSU homogeneity of the sample, and increased design effects. It was decided that the 12 second-stage unit size would be retained, allowing three second-stage units to be allocated to each quarter.

For the 2011-2019 NSFG fieldwork period, a similar review of the number of second-stage units per PSU was conducted. Cost-efficiency (for example, travel cost per completed interview) requires fewer than 12 units. Lower variance estimates generally require more second-stage units. The choice of 12 second-stage units that can be allocated in sets of three across quarters in a calendar year was retained in the NSFG 2011-2019 design because it yields a good balance between cost-efficiency and sampling variance. In large PSUs, a larger number of second-stage units (always a multiple of 12) was selected in order to equalize domain-level sampling rates across all PSUs. Wherever this occurred, multiple interviewers were hired and each interviewer was assigned 12 area segments for the year.

4.1 Second-Stage Selection

The second-stage units (SSUs), termed “area segments,” are Census Blocks or combinations of Census Blocks. Within each sample PSU, segments were implicitly stratified by ordering the list of segments by the density of black and Hispanic households (for example, from high to low, within Block Groups) and systematically selected with probabilities proportionate to composite size measures.

Area segment units were formed within each second-stage high density domain. The first step in this process was to combine neighboring blocks to form area segments that had a minimum

of 50 occupied households (the same method as was used in the NSFG 2006-2010 continuous design). A measure of size was then calculated for each segment. A domain-specific multiplier (see Table 2) was used to assign higher probabilities of selection to segments in high-density minority domains (i.e., domains 2-4). The result of these weighted measures of size is a disproportionate allocation of the area segment selections to high minority domains. This approach yields sampling rates for high density segments that are 2.3 to 2.6 times larger than those for other segments.

The use of weighted measures of size eliminated what was at times a difficult process in the selection of segments. In NSFG 2002 and 2006-2010, an integer number of segments had to be allocated across high density strata within each PSU. The new approach simplified the process of allocating sample over domains and time. The weighted measure of size method eliminated the need for this separate allocation.

In the continuous 2006-2010 NSFG, exactly 12 segments were selected in the non-metropolitan non-self-representing sample PSUs. The 28 larger self-representing PSUs received an allocation of segments that was proportionate to size, with the smallest receiving approximately 12 segments and the largest more than twice as many.

For NSFG 2011-2019, in each calendar quarter within a PSU, one-quarter (3 or 6) of the segments allocated to each PSU in the yearly sample were selected for each 12-week data collection quarter. Each quarter, approximately 114 area segments were released. This led to a total of 456 each year, or 912 area segments in total. During 2015-2017, there were two quarters where the number of segments that were released was reduced in order to reduce data collection costs.

Across the two stages of selection, the probability of selection is $\frac{M_{h\alpha}}{M_h} \times \frac{d_{h\alpha} M_{h\alpha\beta}}{M_{h\alpha}}$, where $d_{h\alpha}$ is the number of segments to be selected in the α^{th} PSU in stratum *h* (usually 12). With the composite measures of size, relatively more high-density segments, and blocks, were selected for housing unit sampling and screening. For PSUs that were in the sample for more than one year, segments were assigned to the quarter using systematic sampling across all of the quarters for years that the PSU was sampled. This list of area segments was sorted by domain and geography.

5. Tertiary Stage: Housing Unit Lists and Sampling

5.1 Preparation of Housing Unit Lists

Once sample segments had been selected, the next step was to prepare lists of housing units for each selected segment. The initial lists came from either a list of addresses for housing units

obtained from a commercial vendor of the U.S. Postal Service's Delivery Sequence File (DSF) or from an original "scratch" listing. All addresses for each segment were requested from a commercial vendor. In the 2015-2017 survey period, less than 1% of segments had no addresses available from the commercial vendor. These low count segments were typically in rural areas where either Post Office boxes or rural delivery addresses are present. For these low address count segments, an interviewer was sent to the segment to list addresses from "scratch." The remaining 99% percent of the segments were deemed eligible for an "update" listing. A stratified sample of these segments was drawn and an interviewer was sent to the selected segments to update the address lists based on what was observed directly.

Over 54% of segments were found to have a high probability of being nearly completely covered by the commercial list of addresses. In turn, a model was developed that predicts whether the DSF is likely to offer complete coverage of all housing units in an area segment. This model was used to stratify the area segments. The area segments that were in the stratum above the selected cutoff value for the estimated probability of complete coverage had a coverage rate of 97%. Among the approximately 50% of segments that had a high probability of being well-covered by the DSF, a sample of segments to be listed was drawn. The sampling rate varied between 40% and 73%. Lines that were added to these sampled segments were given a weight that was the inverse of the probability that the segment was selected for listing. This weight allowed these cases to represent lines that would have been added in segments where no field listing was done. This resulted in moderate cost savings by reducing the amount of listing required, and did not produce additional coverage error since units that were not likely to be on the list were added in a sample of area segments.

The ISR production sampling and field staff have extensive experience with listing for area probability samples. The sampling specialists selected the core sample of area segments. The ISR production sampling group prepared area segment maps using TIGER files and GIS software. Maps guide an interviewer to the exact location of the blocks in the segment. Maps were generated through commercial software using the TIGER files to delineate block boundaries. Three levels of maps were created for each segment: a large-scale view showing the location of the segment relative to major highways and streets, an intermediate scale view showing the segment relative to major streets and block boundaries within the segment, and a detailed small scale view showing individual blocks in each segment.

The maps and the DSF addresses for update segments were loaded onto the ISR Electronic Listing Application (ELA), and released to the interviewer one to two months before the data collection began in the next calendar quarter. Interviewers visited and "scratch" listed (i.e., listed the entire segment from scratch) or updated addresses. Interviewers were trained to update addresses using the software, which allowed them to insert new addresses found in the segment but not on the DSF list, delete DSF addresses that could not be found in the segment, and re-order the list to put the addresses in sequential order around each block to provide further geographic stratification of the sampling and to make the addresses easier to locate after the sample was selected.

The interviewers used the ELA to record street name, house number, apartment number (if applicable), and other location descriptors for each housing unit. They were able to annotate maps (for example, by marking potentially dangerous areas) to clearly describe the location of a unit. The listing proceeded systematically through the segment until all housing units and group quarters in the segment had been listed.

During segment listing, the interviewer also collected and recorded designated observations about the neighborhood as part of monitoring patterns of response and for the contextual data files. This also provided information that the field supervisor and researcher needed to plan their work in the area segment. It included information about potential difficulties such as "dangerous" neighborhoods (high crime areas), locked buildings, and restricted access residential areas. The information was also used to estimate noncontact propensities and guide field follow-up priorities. The availability of this information also alerted field staff to situations that called for the use of escorts, the need to contact law enforcement officers, or the need to communicate with apartment managers before interviewing began.

The update and "scratch" listings were downloaded to ISR as completed. All listings returned from the segment listing program were checked for completeness and accuracy by experienced office staff. The staff used such tools as online maps, satellite photographs, online street views, and original address lists to be sure the list was accurate. A number of automated quality control checks were also used, including comparison of counts of listed units with Census counts of housing units, reporting on violations interviewers had made with respect to consistency checks within the ELA, and outliers in length of time required to complete the listing (based on timestamps from the ELA). Any problems were resolved with the interviewer. Approved listings were transferred to a secure housing unit listing database used for selecting the third-stage sample of housing units.

5.2 Within-Segment Sampling Rates

The selection of within-segment sampling rates took into account the overall design that was targeted to achieve 5,000 interviews per year, with oversamples of females, teens, blacks, and Hispanics; and a labor model that required that there was sufficient work for each interviewer to work an average of 30 hours per week. Under this approach, the sample sizes for each interviewer were allowed to vary such that they would have, in expectation, a large enough sample of housing units in order to work 330-360 hours each quarter.

Beginning in Quarter 13 (i.e., September 2014), a sample design change was implemented to increase the percentage of screened households that contain an eligible person. This was accomplished by stratifying housing units based on a prediction of whether the unit contained an eligible person. The model was selected and estimated using data from previous quarters where the binary eligibility outcome was measured. Key predictors in this model included commercial data that estimate whether an eligible person is in the household. The predicted probability of there being an eligible person in the household was used to create strata and

then oversample the stratum or strata with higher expected eligibility. This approach continued to be used in the 2015-2017 sampling procedures.

The third stage random selection of housing units was made from the segment housing unit list. In order to select addresses and assign them to field data collection, a within-segment sampling rate was set. This rate is a function of the efficiency of the interviewer and, after Quarter 13, the expected eligibility of the housing unit. More efficient interviewers would have relatively higher sample sizes such that every interviewer had enough sampled housing units that they could work 30 hours every week for 12 weeks. Housing units with a higher predicted probability of being eligible would also be selected at higher rates. Once the allocation and listing steps had been completed, a sample of housing units was selected systematically from a geographically-sorted list of housing units beginning from a random start.

The beginning sampling rate for housing units was set to be equal probability selection method (EPSEM) within domain. This rate can be calculated using the following formula:

$$\pi_3 = \frac{\pi_d}{\pi_1 \times \pi_2}$$

Here, π_d is the overall sampling rate for the domain and π_1 and π_2 are the PSU and SSU selection probabilities (described in the previous section). The PSU and Segment notation are suppressed, but each segment has a unique π_3 value. The values for π_d are given in Table 4.

Table 4. Domain level Sampling Rates

Domain	Overall Domain Sampling Rate (π_d)
1	0.000465968
2	0.001211516
3	0.001071726
4	0.001164919

Once these rates (π_3) had been set and the listing of housing units completed, a preliminary expected sample size was calculated. This sample size is the number of housing units listed ($HU_{\alpha\beta t}$) multiplied by the initial rate (π_3) and is denoted ($L_{\alpha\beta t} = \pi_3 \times HU_{\alpha\beta t}$). This preliminary sample size was modified by a multiplier designed to produce a sufficient sample size for a given interviewer efficiency.

The sufficient sample size for an interviewer was calculated at the PSU level. Within an expected 360 hours in a 12-week period, interviewers updated or prepared "scratch" listings for the segments allocated in the next calendar quarter, screened selected lines, and conducted main interviews. Interviewers had in their work assignments varying survey conditions that make them more or less efficient within the 360 hours available. The conditions varied by the nature of the communities in which they worked, which in turn affect parameters such as the number of hours required to complete an interview (i.e., the hours per interview, or for the α^{th}

PSU at calendar quarter *t*, (HPI_{at}); the housing unit occupancy rate (O_{at}); the proportion of occupied housing units with one or more persons ages 15-49 (the eligibility rate, E_{at}); the proportion of the sample that is either completed during Phase 1 or will be retained for Phase 2 (the subsampling rate $\hat{S}_{at}$); and the combined screener and main interview response rate (R_{at}).

Each quarter, the expected number of work hours was based on the labor model specified earlier. The target that interviewers have for their hours each week is 30. This number was usually used in the sample selection equation. Managers monitored interviewers to ensure that they met the target for hours. The sample line assignment process starts from expected hours, say H_{at} for the α^{th} PSU (usually 360 hours per interviewer) at calendar quarter *t.* A unique estimate of the HPI, $\bar{H}PI_{at}$, is generated for each PSU. Estimates for occupancy, eligibility, the subsampling rates, and response rates, $\hat{O}_{at}$, $\hat{E}_{at}$, $\hat{S}_{at}$, and $\hat{R}_{at}$, although denoted at the PSU-level, were actually developed for the sample as a whole. Attempting to estimate these parameters at lower levels (e.g., Census Region) simply led to more variance in the probability of selection weights and did not prove to be accurate. The following formula was estimated for each PSU.

$$A_{at} = \frac{\left(H_{at} / \bar{H}PI_{at}\right)}{\left(\hat{E}_{at}' \ \hat{O}_{at}' \ \hat{S}_{at}' \ \hat{R}_{at}\right)}$$

These housing unit or address sample sizes A_{at} were adjusted after review by study staff to account for interviewer or PSU conditions that depart from expectations for the region and domain. For each PSU α during calendar quarter *t*, the ratio of lines needed for an efficient workload over the lines allocated (summed across all segments in the PSU) under an EPSEM sample of housing units is defined as:

$$D_{at} = \frac{A_{at}}{\sum_{b=1}^{a_a} L_{bt}}$$

This ratio was then used to modify the sample size in each segment for PSU α. Here the notation for PSUs, segments, and time is suppressed for π_3^* and π_3:

$$\pi_3^* = \pi_3 \times D_{\alpha t}$$

Note that this rate might imply a non-integer value number of sampled housing units. Therefore, the probability of selection was not the number of units selected divided by the number of units on the list. The latter rate is close to the actual rate, but may be slightly different because of the need to select an integer number of housing units. Further, during NSFG 2015-2017, the ratio $D_{\alpha t}$ was bounded to be no more than 2.5 and no less than 0.5. The goal was to limit the variation in weights while still allowing sampling rates to be higher for more efficient interviewers and lower for less efficient interviewers.

In a final step, the rates of selection π_3^* were modified by factors F_m designed to produce the desired sampling rates across the housing unit strata denoted l. There were three strata formed from the estimated probabilities of being eligible (predicted low, medium, and high probability of being eligible).

$$\pi_{3m}^{\dagger} = \pi_3^* \times F_m$$

The adjustment factors F_m were set based upon a review of the expected 1+L weighting loss and the expected increase in the eligibility rate under a distribution of options. The stratum with the highest eligibility had the sampling rates for its units raised. The other strata had their sampling rates reduced by a factor that would keep the sample size nearly constant. Given the link between the sampling rates and interviewer productivity, there was a need to implement this change gradually. Therefore, the expected percentage point increase in the eligibility rate increased over time. The sampling rates were set to increase eligibility rates about six percentage points over a design that does not oversample based on predicted eligibility. However, the actual rate varied depending upon the characteristics of the sample and the distribution of the estimated probabilities of being eligible that result. The adjustment factors F_m applied to the sampling rates are included in Table 5.

Table 5. Proportion of Housing Units in Each Stratum, Sampling Rate Adjustment Factor, and Predicted Eligibility by Quarter

Quarter	Predicted Low Eligibility			Predicted Medium Eligibility			Predicted High Eligibility			Expected Percentage Point Increase in Eligibility	Actual Percentage Point Increase in Eligibility
	% In Stratum	Predicted Eligibility	Adj Factor F_m	% In Stratum	Predicted Eligibility	Adj Factor F_m	% In Stratum	Predicted Eligibility	Adj Factor F_m		
Q17	24%	25%	0.78	30%	59%	1.00	46%	86%	1.15	4.2%	4.1%
Q18	16%	17%	0.71	37%	50%	1.00	47%	83%	1.16	4.3%	4.2%
Q19	19%	16%	0.68	27%	46%	1.00	54%	77%	1.18	5.3%	4.9%
Q20	22%	21%	0.69	40%	53%	1.00	38%	82%	1.28	5.5%	6.0%
Q21	21%	23%	0.64	36%	55%	1.00	42%	81%	1.25	5.9%	4.6%
Q22	16%	20%	0.69	32%	50%	0.89	51%	79%	1.25	5.1%	4.8%
Q23	14%	18%	0.67	35%	50%	0.84	50%	80%	1.32	6.0%	5.8%
Q24	15%	18%	0.67	31%	46%	0.79	54%	79%	1.34	6.5%	6.0%

Once the allocation and listing steps had been completed, a sample of housing units was selected systematically from a geographically-sorted list of housing units beginning from a random start using the sampling rates ($\pi_{3m}^{\dagger}$) described in this section.

This allocation leads to variation in probabilities of selection of housing units across segments within and among PSUs. The variation was compensated for through weights (described below), although the added variability in sample weights from varying line probabilities at the segment level has the potential to increase the variability of survey estimates.

5.3 Third-Stage Selection of Housing Units

The third stage random selection of housing units was made from the segment housing unit list. In order to select addresses and assign them to field data collection, a housing unit sampling rate was determined to meet the allocation for the interviewer's sample while also yielding higher eligibility rates as described in the previous section.

Once the allocation and listing steps had been completed, a sample of housing units was selected systematically from the geographically-sorted list of housing units beginning from a random start.

5.4 Screening and Missed Housing Units

The selection of housing units and households was continued in the household screening operation in the field. Screening is the process used to determine whether the selected housing unit is occupied, and then whether any eligible persons reside in the occupied housing unit. Screening consisted of a short questionnaire administered at the doorstep of every housing unit selected for the sample. Using weighted estimates, about 13 percent of units were unoccupied or not actually housing units (for example, a housing unit converted to a commercial building). In addition, among occupied housing units, about 51% contained at least one eligible person.

Table 6 reports the total count of selected addresses, screened eligible households, and main interviews, as well as the averages per quarter of data collection. Data from the 2006-2010 and 2011-2015 NSFG data collections are also presented for comparison.

Table 6. Number of selected addresses, screened eligible households, and main interviews; and average number per quarter, 2006-2010, 2011-2013, 2013-2015, and 2015-17 NSFG.

	2006-2010	2011-2013	2013-2015	2015-2017
Selected addresses[a]				
Total	78,082	39,494	40,598	38,890
Average per quarter	4,880	4,937	5,075	4,861
Screened eligible households[b]				
Total	32,134	15,287	15,239	15,797
Average per quarter	2,008	1,911	1,905	1,975
Main interviews[c]				
Total	22,682	10,416	10,210	10,094
Average per quarter	1,418	1,302	1,276	1,262

[a]Selected (or sampled) addresses are the number of addresses selected into the screener sample.
[b]Screened eligible households are successfully screened addresses containing one or more age-eligible persons.
[c]Main interviews are screened eligible households with a completed interview with the selected respondent (including partial interviews which are those where the respondent at least reached the last applicable question before ACASI).).

Table 7 presents key indicators of eligibility, again with comparable data from 2006-2010 and 2011-2015. Both the percentage of occupied housing units and the percentage of households with eligible persons were lower in 2011-2013 than in 2006-2010. However, both measures rose in 2013-2015. In 2015-2017, the occupancy rate rose again. Between 2013-2015 and 2015-2017, the percentage of households with eligible persons increased from 51% to 58%. This was due to both the expansion of the eligible ages and changes in the sample design – specifically, the oversampling of housing units that are expected to be eligible.

Table 7. Weighted percent of housing units that were occupied, percent of occupied housing units with an age-eligible person, and percent of occupied housing units with access impediments by data collection release, 2006-2010, 2011-2013, 2013-2015, and 2015-2017 NSFG.

	2006-2010	2011-2013	2013-2015	2015-2017
Percent of all housing units that were occupied	85.6%	84.4%	86.3%	87.0%
Percent of all occupied households with an age-eligible person 15-44 (or 15-49 for 2015-2017)	52.3%	48.8%	50.8%	57.9%
Percent of occupied housing units with access impediments*	14.1%	13.6%	15.9%	13.5%

NOTE: Results are based on removal of screener and main lines not selected for the second-phase sample.

*Examples of access impediments include locked apartment building doors and gated communities with guards.

Interviewers had been trained to check for housing units that may have been missed in the update or scratch listing processes. Such missed housing units occur when an interviewer overlooked a structure with one or more housing units, or missed a part of a structure that was a separate housing unit. Missing units may also occur if a housing unit is constructed since the listing took place.

Interviewers were equipped with a sample management system called SurveyTrak on their laptop computers that contained a listing of all housing units in each segment. The procedure for handling missed housing units in the field was as follows: for each sample housing unit designated in the SurveyTrak listing, interviewers checked for full coverage by ensuring that all housing units following the selected housing unit in geographic order were present on the list, and checked for mail boxes, doors, or utility meters that might indicate a unit that was not

listed. They were instructed to ask screener respondents about any additional housing units in the structure. If one or two additional housing units that were not on the list were discovered between the sample housing unit and the next listed unit, the interviewer added them to the SurveyTrak list. The interviewer then attempted a screening interview with the additional units.

When more than two housing units were missed between a selected address and the next address on the list, interviewers were trained to suspend work for that sample address, including contact with the household, and call the ISR's sampling unit to receive further instructions. Before calling, interviewers had been trained to obtain a list of all additional housing units associated with the sample housing unit. The ISR central office staff then subsampled the original and the additional housing units, and returned an updated sample of addresses to the interviewer in the next daily download of the sample. This process created unequal probabilities of selection of housing units within each domain, and so weighting adjustments to account for the missed housing unit subsampling were incorporated into final weights.

6. Household Listing and Respondent Selection

One eligible person per household was selected from all households containing at least one eligible person using a random selection procedure. This random selection of one eligible person reduces measurement error from contamination that may occur if more than one person in a household is interviewed, and reduces the loss of precision due to within-household clustering.

The last stage of sample selection was conducted within the household during the screening activities. An adult member of the household was asked to provide a list of all persons living in the household. Information on the gender, age, and race or ethnicity of each person was recorded during the screening portion of the interview. Interviewers asked additional questions to be sure no one was missed, particularly college students living away from home at a dormitory, fraternity, or sorority. (College students living away from home in their own apartment or housing unit are covered by the household frame, and are not considered to be part of their parents' household.) Dormitory, fraternity, and sorority residents were included in the household listing of their parents' household.

If no one in the household was between the ages of 15 and 49 years, then the screening interview was terminated. If one or more eligible persons were found, the computer-assisted screening system made a selection of one eligible person in the household. That is, one eligible person was selected within each household that contained any eligible persons.

Within-household sampling rates for eligible persons varied by age and gender in order to meet the target sample sizes for teens and females. The selection was made in an application within the Blaise screening instrument. The system was programmed to allow ISR and NCHS staff to

achieve target sample sizes more precisely in the face of uncertainty about rates of eligible persons in the population.

The within-household selection procedure assigns a "measure of size" to each age-eligible person in the household based on the age and sex of the sample person. Larger measures assigned to a subgroup increase the chances that persons in that subgroup would be selected for interviewing (see Tables 8 and 9). Larger measures of size were assigned to teenagers 15-19 years of age to select enough to meet sample size targets. Slightly larger measures were also assigned to females to increase the number of females relative to males in the final sample.

Table 8 (see below) shows the measures of size assigned to each of the four cells created by cross-classifying age and gender for the 2011-2015 and 2015-2017 periods. With the expansion of the age range, more persons within households who had teens were now eligible (i.e., persons 45-49 years of age). Maintaining the proportion of teens required reducing the measures of size for those 20-44. This increased the weights for adults who live with teens. However, in order to mitigate this effect, a smaller target proportion for teen interviews was adopted. Simulation was used to determine the optimal sampling rates. The new measures of size increased the sampling rates for teens relative to adults, but also reduced the expected proportion of teens to 18.2%. The procedure for setting the probability for any person within a household was to divide their measure by the sum of the measures for all eligible persons within the household.

Table 8. NSFG Person-Level Measures of Size

Data Collection Years	Female			Male		
	15-19	20-44	45-49	15-19	20-44	45-49
2011-2015	1.00	0.40	NA	0.93	0.36	NA
2015-2017	1.00	0.25	0.25	0.91	0.23	0.23

Extreme probabilities of selection could have resulted from this algorithm in two situations. The first is if there were a large number of persons within a household. These extreme probabilities of selection would always occur for large households under any sample design where one person per household was selected, although the problem may be magnified by the unequal probabilities assigned for the NSFG. The second situation that resulted in extreme weights occurred when a person with a low measure lived with other persons with larger measures. For example, a 20-49 year old male who lives with three male teens would have ([0.23]/[0.23+3*0.91]=) 0.078 probability of being selected. This would result in a weighting factor of about 12.87 for such a person. Table 9 shows two additional examples of how these measures of size work in practice.

Table 9. Example Family Compositions and Resulting Probabilities of Selection (2015-2017)

Household	Person	Description	Measure of Size	Probability of Selection	Weight
1	1	45-year-old male	0. 23	0.23/(0.23+0.25)=0.48	2.09
	2	40-year-old female	0.25	0.25/(0.23+0.25)=0.52	1.92
2	1	43-year-old female	0.25	0.25/(0.25+1.00+0.91)=0.12	8.64
	2	17-year-old female	1.00	1.00/(0.25+1.00+0.91)=0.46	2.16
	3	15-year-old male	0.91	0.91/(0.25+1.00+0.91)=0.42	2.37

Once each eligible person was assigned a measure of size, the sizes were cumulated, and the total sum of measures recorded. A random number from zero to the total sum of the measures in the household was generated by the sample screener application. The first listed person whose cumulative measure of size within the household exceeds the random number was selected. The chance of selection of the person was proportionate to the relative size of their measure of size in the household.

7. Two-Phase Sampling

Each NSFG 2015-2017 calendar quarter consisted of two phases. In the first 10 weeks of the quarter, interviewers screened selected lines in assigned segments, conducted main interviews in households with eligible persons, and updated or prepared "scratch" listings for the segments allocated in the next calendar quarter.

After 10 weeks of data collection, there remained addresses that had not been successfully screened and sample persons who had not yet completed the interview. If the data collection were halted at the end of 10 weeks, these unscreened lines and persons who had not been interviewed could have contributed to nonresponse bias. A "double" or "two-phase" sample design (Hansen and Hurwitz, 1946) was instituted for the remaining two weeks of the quarter as a strategy reduce the nonresponse bias in survey statistics. The approach was expanded by Groves and Heeringa (2006) to say that the design across each phase should be complementary such that the biases across the phases cancel each other out.

There were two impacts of a two-phase design. First, if the second phase protocol is successful in measuring 100% of the *sampled* nonrespondents from the first phase, nonresponse bias would be eliminated. In practice, no subsample of nonrespondents attains a 100% response rate and thus some nonresponse bias remains, but the bias is expected to be reduced by the

capture of data from the first phase nonrespondents. Second, the cases sampled into the second phase that are successfully interviewed are assigned new selection weights (reflecting the fact that they must "represent" the nonselected nonrespondents). This additional weight component generally increases the variance of the estimates.

Two-phase designs are increasingly attractive to survey researchers because they offer a way to control the costs at the end of a data collection period while addressing concerns about nonresponse rates and errors. In face-to-face surveys, at the end of the data collection period, large costs are incurred for travel to sample segments to visit only one or two sample units, usually those extremely difficult to contact in prior visits or repeatedly displaying some reluctance to grant the survey request. By restricting these expensive visits to a sample of the nonrespondents at the end of the study, this method limits costs while addressing the need to increase response rates and reduce nonresponse bias.

In the 2015-2017 NSFG, a subsample of nonrespondents was chosen for weeks 11 and 12 of each quarter based on review of the history of the first 10 weeks' sample. Study staff developed response propensity models to predict the probability that a given case would yield a completed screening interview or a completed main interview (see Groves et al., 2009, for details of the propensity models). Within a PSU, two of the three segments were sampled at random. The probabilities of selecting a segment were proportional to the size using the sum of the estimated response propensities as the measure of size. The active nonresponse cases in the two remaining segments were grouped into four strata at the conclusion of the 10-week Phase 1 data collection. The cases were first categorized as unscreened or identified eligible sampled persons. Among the unscreened cases, those that were predicted to be eligible (based upon a logistic regression model including paradata and sampling frame data used in response propensity models, supplemented with information from commercial databases regarding the ages of persons within unscreened households) were reclassified as "identified eligible person." Within each of these groups, cases were classified as medium-high or low propensity to respond, based on tertiles of the estimated response propensities. This created a 2 x 2 classification of all active cases. A disproportionately allocated sample of nonresponding cases was selected across these groups or second phase strata, with higher probabilities of selection from strata with higher likelihood of response and from strata with known or predicted "eligible persons." These selected lines and persons were then released to interviewers for Phase 2 data collection in the last two weeks of the calendar quarter. Table 10 presents the counts of unscreened units and identified eligible sampled persons selected for Phase 2 by quarter for NSFG 2015-2017.

Table 10. NSFG 2015-2017 Second Phase Samples by Screening Status and Quarter

Quarter	**Unscreened Units**	**Identified Eligible Persons**
17	257	263
18	274	192
19	296	199
20	219	227
21	264	252
22	245	246
23	242	200
24	326	262
Total	2,123	1,841

Flowing from the responsive design perspective (Groves and Heeringa, 2006) that guides the NSFG's design and fieldwork, study staff implemented a Phase 2 interview recruitment protocol that was distinctive from that used in Phase 1. Such a distinction is necessary (but not *a priori* sufficient) to attract sample persons who did not find the Phase 1 protocol effective, and thus increase response rates and reduce bias in the sample data. Peytchev, Peytcheva and Groves (2010) provide evidence that the second phase protocol used in NSFG Cycle 6 was effective at bringing in persons for whom the Phase 1 protocol was ineffective. With the approval of two Institutional Review Boards and the Office of Management and Budget, the Phase 2 recruitment protocol in the 2015-2017 NSFG involved the following components:

a. a prepaid $5 token of appreciation payment (versus none) for cases that had not yet completed the screening interview;
b. a prepaid $40 token of appreciation for the main interview; and
c. a promised additional $40 token of appreciation for a completed main interview.

8. Sample Selection in a Responsive Design Framework

Surveys with high response rate goals and limited budgets such as the NSFG need to be able to monitor key survey design parameters such as completed interviews, eligibility rates, response rates, expenditures, and interviewer productivity. In most surveys, the information systems which provide such data are designed to provide some data daily and other data at the end of the data collection period. There is seldom an opportunity to make changes to a survey design based on this information between the start and end of data collection.

In the 2015-2017 NSFG, paradata monitoring techniques allowed a level of monitoring such that it was possible to make survey design changes at any point during the data collection. The information systems provided daily data on how many interviewer hours were being used in data collection, what areas and interviewers had good and poor results, and what types of nonresponse was most prevalent in an area.

NSFG project staff used the NSFG Dashboard and other information systems to manage data collection to keep within budget and meet survey sample size and data quality targets. Continuous interviewing integrated well with management interventions based on data from the information systems that are used regularly to adjust data collection to ongoing survey conditions.

As described in the previous sections, the 2015-2017 NSFG's continuous interviewing design relied on four sampling levels: the PSU, the segment, the housing unit, and the person within the housing unit. A fifth level was added to allow management to achieve response rate goals: a second phase (or "double") sample for nonresponse.

The sample levels permitted design changes made throughout the field period that were aimed at adjusting sample sizes based on eligibility rates, response rates, and interviewer performance. These management manipulations constitute design elements of an NSFG continuous survey "responsive design" (Groves and Heeringa, 2006) that are used to control sample size and response rates.

The first level of the responsive design process was setting sample size at the PSU level. A quarterly review of SurveyTrak data was used to ensure that sample size targets would be met. These sample size targets for each PSU were projected before each annual sample began from the most recent data available on expected interviewer workload and performance in the same PSU or a similar PSU, PSU-specific eligibility rates, and past or estimated PSU-level response rates.

Within an annual sample, there was a further opportunity for responsive design at a second level, the segment selection. As described above in more detail, Census Blocks within PSUs were divided into density domains on the basis of the concentration of black and Hispanic populations within the corresponding block group. A single sample of segments was selected across all domains simultaneously to allow 12 or more segments to be selected in most PSUs. Each set of segments within a PSU was divided further into four sets for each of the 12-week quarters to be released approximately in September, January, March or June in each data collection year. The number of segments and size of sample from each segment in each quarter could be adjusted based on SurveyTrak data to reflect interviewer workload and performance, expected eligibility rates, and expected response rates. If, for example, the target sample sizes needed to be cut, the ISR design allows reduction in the number of segments within PSUs, or the reduction of the housing unit sample sizes selected within the segments.

A third level of sample selection was the housing unit sample within segments. The sample selection rates and cluster sizes could be varied across segments depending on the housing unit yield of the listing operation in order to yield a number of housing units in each interviewer's assignment for the calendar quarter to match interviewer efficiency (hours per interview, described below) as well as expected response, occupancy, and eligibility rates. Further, if sample sizes for race or ethnicity groups were projected to fall below targets, within-segment rates could be increased for high-density black and Hispanic population segments, responding to existing survey conditions.

The fourth level of selection was the random choice of a person ages 15-49 within the household. Interviewers visit selected housing units in assigned segments starting at the beginning of the calendar quarter. A household roster was generated containing a list of all persons who usually reside in the household. The within-household selection probabilities could be varied from one calendar quarter sample to the next to adjust sample size to achieve target sizes for teens and females.

A fifth level of selection was introduced in each calendar quarter that provided an opportunity to respond to response rate and sample size yield. Each 12-week calendar quarter was divided into a 10-week Phase 1 sample, and a two-week Phase 2 sample. At the end of the 10-week Phase 1 sample, selected addresses remained that had not been successfully screened or had been successfully screened but did not have a main interview completed. A sample was drawn of these outstanding addresses, typically about one-third of the total within a PSU. This second phase sample was selected for Phase 2 sample interviewing during the last two weeks of the calendar quarter. Interviewer assignments were reduced so that interviewers concentrated effort on a smaller number of addresses and sampled persons for the final two weeks of data collection. This within-quarter selection was particularly useful to control final response rates and costs for the overall sample in the calendar quarter.

9. Weighting to Compensate for Unequal Probabilities of Selection

A base or starting strategy with most survey sample designs is to consider a representative sample, one that is a "scale model" of the population from which the sample is to be selected. However, smaller groups in the population may have too few cases in the sample to provide adequate precision for those groups. Survey sample designs such as the NSFG thus deliberately over- and under-represent smaller groups in the sample. This over- and under-representation is accomplished through the use of varying probabilities of selection. Over-represented groups have higher sampling rates than under-represented groups.

For example, non-Hispanic black women represent approximately 7.5 percent of the population 15-49 years of age as of 2015. Yet, for purposes of improved precision for non-Hispanic black women, NSFG 2015-2017 chose the sample in such a way that these women account for about 13.4 percent of all respondents in the sample. Similar kinds of over-representation have occurred for non-Hispanic black men, Hispanic women and men, and teenagers of all races. Of

course, the over-representation of these groups means than non-Hispanic, non-black men and women ages 20-49 are under-represented in the samples.

These kinds of over-sampling rates were used for the 2015-2017 NSFG. As in previous NSFG surveys, "sampling weights" were needed to adjust for these different rates and this over-representation. Without appropriate weighting, resulting estimates of fertility and family growth characteristics could be subject to substantial bias.

In the sample design described in the previous sections, the over- and under-sampling design is more complicated than disproportionate representation by race/ethnicity, age, and sex. In order to implement the over- and under-sampling, a respondent's sampling rate was, in the sample design, determined by the response rate, eligibility rate, occupancy rate, and efficiency of data collection in a PSU; the segment domain (persons living in block groups with more black or Hispanic persons have higher sampling rates); the age, sex, and race/ethnicity of the individual relative to others within the same household; and the second phase subsampling which depends on a complex set of factors used to predict the daily response propensity. All of these factors must be taken into account when developing an appropriate weight for the purposes of compensating for over- and under-sampling in the sample selection.

The sampling weights for the design were comprised of three components: an adjustment for unequal probability of selection, a unit nonresponse propensity adjustment, and a post-stratification factor. The adjustment for unequal probability of selection is discussed in this section, since it is most closely related to the sample design described in the previous sections. The procedures to develop the latter two components of the final sampling weights are described in the following sections.

For purposes of description, it may be useful to observe that the sampling weight can be interpreted as the number of persons in the population that an individual respondent represents. A final sampling weight for a teenage non-Hispanic female of 2,000 means that that sample respondent represents herself and 1,999 other similar women in the population. The 2015-2017 NSFG final weights are values greater than 1, and when summed across a subgroup, or the total sample, are expected to provide an estimate of the total number of persons in the subgroup in the population.

9.1 Inverse Probability Selection Weighting

Each stage of selection must be included in the development of selection weights. The description of these weighting factors follows the order of the sample selection.

Each of the 65 PSUs included in the 2015-2017 NSFG was selected with probabilities proportionate to a composite of the number of occupied housing units across the four within-PSU sampling domains. The probabilities were computed at the time of selection and stored for each sample line.

Census Blocks were selected within PSUs with probabilities proportionate to composite measures of size based on Census counts of the number of occupied housing units reported in the Census 2010 data (using the Redistricting data in year 1 and Summary File 1 in subsequent years). In all PSUs, segments in minority (high density black, Hispanic, and both) domains are chosen with higher probabilities of selection than for those in the non-minority domain. Probabilities of selection for segments within domains were proportionate to the estimated number of households in the segment. The probability of selection of each block was computed at the time of selection and stored for later use.

Not all housing units in sample segments were selected for the NSFG sample. Housing units within area segments were sampled in order to achieve a target number of sample housing units or lines. The subsampling probabilities of selection for housing units varied by segment and were stored with the record for each unit.

Once sample housing units or lines had been selected and released through the ISR sample management system to interviewers in each PSU, the interviewers visited sample housing units to determine if any eligible persons resided there. The interviewer completed a household roster in the Blaise instrument recording age, gender, and race and ethnicity for each member of the household. The age eligibility for NSFG 2015-2017 was 15-49 years of age. If one or more age-eligible persons lived in the sample housing unit, a random selection was made of one eligible person per household with chances of selection varied to increase the selection of teens (ages 15-19) and women. The household roster and chances of selection were recorded in the Blaise household record.

Finally, Phase 2 sampling in the final two weeks of the calendar quarter data collection varied the selection probability of those cases that had not been completed by the end of Phase 1. On average, a sample of two out of three segments in a calendar quarter was chosen in each PSU. Nonresponding housing units in the selected segments (those which have not reached a final disposition after 10 weeks of data collection) were divided into strata on the basis of type (screener or "main"—the latter including cases judged likely to be eligible) and predicted probabilities of obtaining a completed interview. A Phase 2 selection of segments and of nonresponding housing units was chosen, with higher chances of selection assigned to those nonresponding housing units with higher estimated response propensities and higher chances of selection for eligible or likely eligible cases. The varying chance of selection for segment and housing units in the second phase selection was retained for subsequent weighting.

9.2 Probability of Selection and Weight

The probability of selection of each sample person can thus be computed using the probabilities of selection for PSUs, segments, sample housing unit, within household selection, and Phase 2 subsampling of active cases.

Let $M_{h\alpha}$ denote the composite size measure of the α^{th} PSU in PSU stratum *h,* the number of occupied households in the PSU as measured in the 2005-2009 American Community Survey Summary File, modified by a weighting factor (see Table 2) to account for over-sampling in minority domains. Let $M_{h\alpha\beta}$ denote the composite size measure for the β^{th} segment in the $(h\alpha)^{th}$ PSU, where again the size measure for each segment is the number of occupied housing units in the 2010 Census increased to account for over-sampling in minority domains and $d_{h\alpha}$ is the number of SSUs selected in the $(h\alpha)^{th}$ PSU. Also, let $\pi^{*}_{3,h\alpha}$ denote the sampling rate calculated to obtain the desired number of sample lines in the $(h\alpha)^{th}$ PSU as described earlier. Finally, let $\pi_{w,h\alpha\beta\gamma\delta}$ denote the within household probability of selecting the $(\delta)^{th}$ person within the $(h\alpha\beta\gamma)^{th}$ household in a segment where γ denotes the housing unit within the $(h\alpha\beta)^{th}$ segment, $\pi_{2,h\alpha\beta}$ denote the Phase 2 selection probability for the $(h\alpha\beta)^{th}$ sample segment, and $\pi_{2,h\alpha\beta\gamma}$ denote the Phase 2 selection probability for the $(h\alpha\beta\gamma)^{th}$ household within the selected Phase 2 segment.

The probability of selection of the $(h\alpha\beta\gamma\delta)^{th}$ eligible person is computed as

$$\pi_{h\alpha\beta\gamma} = \left(\frac{M_{h\alpha}}{\sum_{\alpha=1}^{a_h} M_{h\alpha}}\right) \times \left(\frac{d_{h\alpha} M_{h\alpha\beta}}{\sum_{\beta=1}^{b_{h\alpha}} M_{h\alpha\beta}}\right) \times \pi_3^{*} \times \pi_{w,h\alpha\beta\gamma\delta} \times \pi_{2,h\alpha\beta} \times \pi_{2,h\alpha\beta\gamma}$$

The base weight compensating for unequal chances of selection for the $(h\alpha\beta\gamma\delta)^{th}$ eligible person is the inverse of this probability of selection, $w_{h\alpha\beta\gamma\delta} = w_{1i} = \pi^{-1}_{h\alpha\beta\gamma\delta}$.

A further stage of sampling occurs when subsamples of PSUs and SSUs are taken over time. For example, a subsample of 24/192 nonself-representing PSUs were randomly assigned to each year of data collection. The 2015-2017 NSFG is one such subsample of the full NSFG 2011-2019 sample design. See the Weighting Documentation for additional details.

The base weights vary across persons due to the over- and under-sampling within households. The variation has the potential to increase the variance of estimates. The full effects of these increases are reduced later in the weighting process through trimming of the largest final weight values. These weights were calculated at the time of selection and were monitored for extreme outliers using statistical process control techniques.

10.Post-Survey Adjustment

10.1 Post-Survey Adjustments for Unit Nonresponse

For the NSFG 2015-2017 public-use files, both sample-based weighting adjustments and population-based (post-stratification) adjustments were used to reduce error from unit nonresponse. Unit nonresponse is the failure to obtain data for a selected unit by the end of data collection activities. Although unit nonresponse adjustment can use a single final response status as the outcome (for example, interviewed or not interviewed), this practice ignores the important distinction between the two stages of interviewing in the NSFG: screening for eligible persons and gaining cooperation for the main interview. The screening interview is relatively low burden and may be conducted with any adult in the household. The main interview has a higher burden and is conducted with a specific person who has been randomly sampled. Further, the data available on these two stages of interviewing differ. Data from the screening interview, including interviewer observations about the sampled person, can be used to examine whether responders and nonresponders differ. For the nonresponse weighting for NSFG 2015-2017, separate adjustments were used to account for the different influences at each stage in order to improve the potential nonresponse bias reduction that can be achieved through unit nonresponse weight adjustment.

In addition, survey statisticians advocate the use of two kinds of data as nonresponse predictors in the adjustment process. One is paradata collected routinely throughout the data collection process, such as contact observations and call records. The other is deliberate interviewer observation on a limited set of potential weighting adjustment predictors that can be used to develop models more predictive of survey cooperation processes and, simultaneously, the survey data themselves. NSFG 2006-2010, for example, used both of these kinds of data in the unit nonresponse adjustment process (see Lepkowski, et al., 2013). The current data collection protocol includes paradata collection at the listing, the calling, the contact, and the interviewing phases of NSFG, and further development of the collection of interviewer observations on factors thought to be related to nonresponse and to the underlying fertility and family growth data collected in NSFG. Although paradata regarding the level of effort applied to each case may be strongly related to response, they are often only weakly related to survey data collected in the main interview (Wagner, et al., 2014). In order to be included in nonresponse adjustment models, the potential predictors needed to have correlations with the main interview data.

Unit nonresponse has and will continue to occur in NSFG Continuous 2011-2019 at two levels: screening to identify sample eligible persons in sample households and main interviewing among selected eligible persons. There is also nonresponse at the initial contact level in the screener interviewing process, but there is expected to be so little data available for non-contact addresses, and so little nonresponse at the contact level, that unit nonresponse adjustment is not feasible. In the following, contact nonresponse is part of the screener nonresponse.

"Sample-based" unit nonresponse adjustments have been used to generate predicted probabilities of response using all available data for respondents and nonrespondents at the screener and main interview levels. As noted above, screener and main interview cases have different cooperation processes that are modeled separately in the adjustment process. In addition, there are slightly different data available at each level. Main interview nonresponse can occur at any time after the conclusion of screening – that is, after a sample person had been selected. The main interview response and nonresponse cases therefore have household composition with race or ethnicity, age, and sex for all persons in the household. A two-step screener followed by main nonresponse adjustment affords the use of a broad range of sampling frame data and paradata at the screener level adjustment, and the same data plus household composition data at the main interview nonresponse adjustment.

This nonresponse adjustment for the NSFG implements an assumption widely used in the adjustment of survey data – the missing at random (MAR) assumption. One expression of this assumption is in terms of classes or subgroups of a sample where, within subgroups of housing units in the screener or selected persons in the main interview, it is assumed that the nonrespondents are a random sample of all the units in the subgroup. A nonresponse weighting adjustment developed under this assumption is computed as the inverse of an estimated response rate or propensity within a subgroup. This is a sample-based weight adjustment that, under the MAR assumption, substitutes an estimated response propensity for a probability of selection in the response process – the probability that a unit will participate in the survey. Thus, as for unequal probability weighting, the inverse of the predicted probability of response serves as an adjustment factor.

There are a number of ways to estimate response rates (see, for example, Groves and Couper, 1998, or American Association for Public Opinion Research, 2015). The methods can be thought of as simply alternative ways of estimating probabilities of responding under MAR. For example, a weighting class method divides the sample into subgroups called weighting classes across which response rates and the characteristics of sample persons are expected to vary. The weighting class method includes a cross-classification of variables used to form the weighting class, and can require large sample sizes overall in order to achieve large enough samples within all subgroups that response rates can be estimated with reasonable reliability. An alternative is the use of a logistic or probit regression model to estimate probabilities of response through a propensity model. The 2015-2017 NSFG, following the approach used in the 2011-2013 NSFG and 2013-2015 NSFG, used this latter method, computing separate logistic models for screening and main interview levels of nonresponse. Different sets of predictors were used for each of the two levels in order to, as indicated above, provide models that took into account the different response processes operating at each level.

The following model development, estimation, and adjustment process was used to develop nonresponse adjustments for the 2015-2017 public-use file (see the weighting document for additional details). Let $S_i = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$ be a zero-one indicator variable denoting whether a sample address has been successfully screened to determine whether eligible persons lived in the

household. The value 1 denotes successful screening and 0 denotes non-contact as well as addresses where screening interviews were refused or not completed for other reasons. This indicator S_i is not defined for unoccupied sample addresses. The screener level logistic regression model is $\pi_{(s)i} = \Pr(S_i = 1 | X_{(s)i}) = (1 + \exp(-X'_{(s)i}\beta_s))^{-1}$ where $X_{(s)i}$ is a vector of predictor values for the *i*th occupied housing unit and $\beta_{(s)}$ is a vector of coefficients. Standard maximum likelihood estimation was used to obtain estimated coefficient values $\hat{\beta}_{(s)}$. These in turn were used to predict the probability of screener completion propensity $\hat{\pi}_{(s)i} = \exp\left(-\hat{\lambda}_{(s)i}\right) \Big/ \left(1 + \exp\left(-\hat{\lambda}_{(s)i}\right)\right)$, where $\hat{\lambda}_{(s)i} = X'_{(s)i}\hat{\beta}_s$ is the predicted logit.

At the main interviewing level of adjustment, $M_i = \begin{Bmatrix} 1 \\ 0 \end{Bmatrix}$ denotes another zero-one main interviewing indicator for the *i*th successfully screened occupied housing unit. M_i is thus 1 when a selected eligible person has a completed interview, and 0 otherwise. M_i would be missing, or undefined, for all sample addresses that are not occupied or a completed screener was not obtained. The main interview logistic regression model is then $\pi_{(m)i} = \Pr(M_i = 1 | S_i = 1, X_{(s+m)i}) = (1 + \exp(-X'_{(s+m)i}\beta_{s+m}))^{-1}$ where $X_{(s+m)i}$ is a vector of predictor values for the *i*th selected eligible person that includes screener as well as household roster data obtained prior to the main interview. Here, $\beta_{(s+m)}$ is a vector of coefficients. Maximum likelihood estimation methods were used to generate $\hat{\beta}_{(s+m)}$ and predicted logits $\hat{\lambda}_{(m)i} = X'_{s+m}\beta_{s+m}$ From the predicted logits, the predicted probability of main interviewing were calculated as $\hat{\pi}_{(m)i} = \exp\left(-\hat{\lambda}_{(m)i}\right) \Big/ \left(1 + \exp\left(-\hat{\lambda}_{(m)i}\right)\right)$.

The main interviewing unit nonresponse adjustment is thus conditional on having completed a screener interview. These estimated screener and main response propensities were used to create nonresponse weighting adjustments. Below is a description of how the model predictors were selected and how the weighting adjustments were created.

The predictors in $X_{(s)i}$ and $X_{(s+m)i}$ were a set of variables that overlap with previous iterations of the NSFG, including 2002, 2006-2010, 2011-2013, and 2013-2015. The potential predictors available at both the screener and main interview levels include: 1) counts and rates for the segment from which the housing unit is selected derived from 2010 Census data for the Blocks in the segment or American Community Survey data from the Block Group or Tract that contains the segment; 2) data obtained from observations made at the segment and housing unit recorded by the interviewer during the segment updating or scratch listing; 3) contact behavior and statements recorded by the interviewer at each contact with anyone within the housing unit; 4) operational measures, such as number of calls to a housing unit, number of

calls to the sample person, and interviewer response rate derived from available paradata; 5) for the main interview propensity model, data derived from the household roster and other data collected in the screening interview; and 6) a limited set of interviewer judgments made at the screener or main interviewing level that are of characteristics related to response propensity and related to fertility and family growth (for example, whether at the screener level the interviewer believes there is anyone under age 15 in the household). ISR researchers have been investigating the utility of these measures (Kreuter et al., 2010), including methods for improving them (West, 2010). These "tailored" adjustment variables provide the best prospect for reducing bias and, possibly, variance (Little and Vartivarian, 2005). Commercial data appear to have less utility for nonresponse adjustment purposes (West, et al., 2015). However, some commercial data have been included based on correlations with survey data and response.

There are a large number of variables in these sets of variables, and not all can be used in a response propensity model, whether at the screener or the main interview level. Variables were identified that are correlated with both response and the survey outcome variables. A two-step process was used to identify the predictors in the nonresponse adjustment models. First, available variables were examined to determine their association, at least among respondents, with fertility and family growth measures. Then a subset of variables with associations to the survey data entered a stepwise logistic regression model predicting unit nonresponse for both screener and main interviewing in the order of their descending correlation with key estimates from the NSFG. This is a form of variable selection for models with limited dependent variables where the value of the dependent variable is not observed (in this case, among the nonrespondents).

Potential predictors in the screener propensity models included such variables as whether housing units in sample segments are residential, whether the address is in a structure with multiple housing units, whether informants at an address asked one or more questions or made statements indicating reluctance to be interviewed during one or more contacts, the total number of calls made to the address, the percentage of calls made during evening hours, whether the segment where the address is located has a higher proportion of population that is black or Hispanic, and whether in the interviewer's judgment there is a person under the age of 15 living in the household. This is only a short selection from a much longer list of predictors that were examined at the screener level.

The main interview response propensity model development had more predictor variables available than the screener propensity model derived from the household roster used in sample selection in the screener. Variables such as age, sex, race, and ethnicity of the selected individual as well as the age, sex, race, or ethnicity composition of the household could be used along with interviewer subjective assessment of several household or personal characteristics made at the first contact with the sample housing unit. For example, interviewers assessed whether there is an active sexual relationship in the household (i.e., a married or cohabiting couple, as indicated by a screener informant referring to "my husband" or "my partner"). This assessment has proven to be correlated with key fertility and family growth variables, and thus may be useful in a predictive model.

The use of the inverse of predicted probabilities as unit nonresponse adjustment weights can lead to substantial variation in response propensity weights. A common practice in survey estimation is to reduce this variation by grouping predicted values into classes, and then using a middle value (often the mean or median) to represent the entire group's predicted values. Since the propensities are estimates, this approach is also more robust to model specification and estimation error. For example, the predicted probabilities could be grouped by deciles of the predicted probabilities. All completed screener cases in the lowest predicted probability decile could be assigned the value of the mean of the predicted probabilities in the decile. The second decile can then be examined, and a similar procedure employed to reduce variation in the weights. The process can be applied to all 10 deciles, or a subset of the deciles can be reduced in this way. This was the approach used to develop nonresponse adjustments for the 2015-2017 public-use file.

The final step in the construction of the nonresponse adjustment weight, the distribution of the weights $w_{(r)i}$ was examined for outlying values. Extremely large values of the weight do not occur with the response propensity class method of trimming described above. However, the impact of larger weights on the variability of the entire distribution of weights was examined to determine if additional trimming, in the form of larger response propensity classes was needed.

Finally, the last step in the construction of the nonresponse adjustment weight is the adjustment of the base sampling weight by the unit nonresponse adjustment weight:

$w_{2i} = w_{si} \times w_{(r)i}$, where W_{Si} is the base sampling weight.

10.2 Post-stratification

Post-stratification is a population-based weighting adjustment. Post-stratification adjustment reduces variances through external population totals for ratio adjustments. These adjustments may also reduce bias for noncoverage and nonresponse. Post-stratification has been consistently applied at the last stage of weighting adjustments in the NSFG since Cycle 1 (1973).

Post-stratification is limited to a set of respondent variables on which population estimates are available. Post-stratification by age, gender, and race/ethnicity, is common because of the availability of population estimates of the sizes of those subpopulations from the Census Bureau. Let W_g denote the proportion of the population in the g-th subpopulation and w_g denote the corresponding sum of fifth-step nonresponse adjusted weights for interviewed persons. The simple post-stratification adjustment is the ratio W_g / w_g for each cell.

The 2015-2017 NSFG continued the practice of post-stratification adjustment to external population estimates by age, sex, and race and ethnicity. This was accomplished by adjusting the nonresponse-adjusted probability of selection weight W_{2i} to an estimate of the civilian non-institutionalized population in the group provided by the Census Bureau. Since the target population continues to include military personnel living off base, the Census Bureau

population projections included such personnel. For the 2015-2017 public-use file, the Census Bureau provided the combined data for non-institutionalized population and military personnel living off base. Across age, sex, and race-ethnicity cells, a post-stratification weight W_{3i} was computed as the ratio of the combined Census and military counts to the sum of the nonresponse adjusted weight W_{2i} in each cell. A preliminary final weight $W_{4i} = W_{3i} \times W_{2i}$ was computed for each of the completed interview cases.

10.3 Weight Trimming

Extreme variation in weights can inflate the variance of survey estimates. Often, it is the case that the most extreme weights can inflate the variance without changing the estimates. In this situation, the extreme weights only inflate the total mean squared error. Trimming these weights is a common practice for surveys in order to reduce the estimated variance without increasing any nonresponse bias. Considerable reduction of the variability of the weights can be achieved by a reduction of a few extremely large weights. The variation of the weights can be examined using the formulation $1+L = \frac{n\sum_{i\in r} w_{4i}^2}{\left(\sum_{i\in r} w_{4i}\right)^2}$ (Kish, 1992), where the elements are the respondents. This 1+L factor approximates the inflation of the variance due to weighting under the assumption that the weights are random. This assumption is conservative as the weights are likely to have some correlation with survey variables. As such, these factors represent a worst-case scenario of the impact of the weights on variance estimates.

In addition to post-survey weight trimming, several steps were implemented for the sampling procedures themselves that reduced this variation prior to any trimming. Further, components of the weights were trimmed by smoothing over differences in weights for cases with similar values on demographics and key statistics (see the weighting documentation for more details).

The weight trimming process took the following steps (see the Weighting Document for additional details). First, the variation in the weights was examined. Outlying weights at both ends of the distribution (i.e., very small and very large weights) were identified. The impact on estimates of trimming the tails of this distribution was then examined. The trimming included taking the sum of the trimmed weights within each post-stratification cell, and redistributing it proportionately across the cases that were not trimmed within the same cell. This was done iteratively until no weight was above the specified minimum or maximum value for the weights. This had the effect of maintaining the post-stratification after the trimming step was complete. This step was completed with different levels of trimming. For each level of trimming, the impact on point estimates and variances across several key statistics was evaluated. Trials of trimming of the following percentiles were made: The 1st and 99th percentiles, 2nd and 98th, 3rd and 97th, 5th and 95th and 10th and 90th percentiles. The trimmed weights were then used to estimate the 19 key statistics (10 for females and 9 for males). The criterion for selecting which weights to trim could be reduction in Root Mean Squared Error (RMSE). However, since this

was evaluated only for a sample of estimates, a somewhat conservative level of trimming was chosen, rather than risking introducing bias into estimates that were not part of the sample. The decision was made to trim the weights at the 1st and 99th percentiles.

After trimming the extreme weights, the cases that had not been trimmed had their weights increased such that the sum of the weights within each cell was still equal to the population control total. If any weight was increased above the specified level for trimming the weights, the trimming and re-post-stratification steps were repeated until no weight exceeded the specified limits. Table 10 lists the sample size, mean weights, and 1+L factors for the overall sample and for several key subgroups for the 2011-2013, 2013-2015, and the 2015-2017 NSFG data collections.

Table 11. Mean final weights (after post-stratification to Census data and trimming), and potential increases in variance due to the weights (1 + L), by sex, age group, and race/ethnicity, 2011-2013, 2013-2015, and 2015-2017 NSFG.

	2011-2013			2013-2015			2015-2017		
	Sample size	Mean weight	Increase in variance (1+*L*)	Sample size	Mean weight	Increase in variance (1+*L*)	Sample size	Mean weight	Increase in variance (1+*L*)
Total	10416	11655.8	2.27	10210	12018.5	2.03	10.094	14249.6	2.37
Male	4815	12569.0	2.19	4507	13572.4	1.93	4,540	15774.7	2.24
Female	5601	10870.8	2.34	5703	10789.9	2.10	5,554	13002.9	2.48
15 to 19	2131	9272.8	2.24	2027	9633.6	1.89	1,821	10746.6	2.26
20 to 49*	8285	12268.8	2.25	8183	12609.6	2.03	8,273	15020.7	2.35
Hispanic	2495	9370.5	2.36	2259	10597.0	2.10	2,060	13334.4	2.44
Black	2192	7803.9	2.41	2069	8305.0	2.07	2,284	8806.2	2.79
Other	5729	14125.0	2.08	5879	13870.8	1.92	5,731	16752.9	2.16

* 20 to 44 for 2011-2013 and 2013-2015.

10.4 Variance Estimation

An important measure of the quality of an estimated statistic (such as a proportion or a mean) from a complex sample survey like NSFG is sampling variance or sampling error (the two terms are often used interchangeably). The sampling variances are used in the computation of confidence intervals, for example, that express the quality of estimates by presenting a range of values to represent underlying uncertainty in the estimate computed from a sample, rather than from a complete census of the population ages 15-49 years. In NSFG 2015-2017, the sampling variance measures variation caused by interviewing a sample instead of the total population of women and men 15-49 across the country, which is more than 143 million. The size of the sampling variance, and derived standard errors and confidence intervals, is determined in part by the sampling design. Features such as sampling weights, stratification, and clustered sample selection must be incorporated into estimates of sampling variance to

obtain good measures of quality of survey estimates. Many statistical software systems have options to estimate sampling variance accurately from a complex sample such as the NSFG, and thus estimation of sampling variance taking into account weights, strata, and clusters in complex surveys is increasingly common and simple. For example, SAS can estimate sampling variances through Taylor series expansion or repeated replication procedures for complex survey estimates in specialized "SURVEY" procedures. Similarly, Stata has Taylor series procedures for complex surveys in "svy" commands as well.

Sampling Error Computing Units. The NSFG 2015-2017 sample has 65 PSUs. Each sample PSU was drawn from a stratum of one or more PSUs. Sampling error computation requires that there be at least two units, or PSUs, in each stratum; therefore, sampling error cannot be estimated directly from this type of a design, which often had a single PSU selected per stratum. In order to allow the estimation of variance, and to prevent disclosure of the identity of PSUs, a set of sampling error or pseudo- strata and clusters were formed from the actual clusters and strata used to select the sample. However, given the need to represent areas as well as years, there were four units in each pseudo-stratum. Therefore, the NSFG 2015-2017 PSUs were arranged into a set of sampling error computing units (SECUs) that in turn were grouped into strata with four SECUs each for variance estimation purposes.

For purposes of variance estimation, the sample selected from each self-representing (SR) PSU (see above) was divided into an even number of representative units by taking a set of interpenetrated systematic samples of the segments within the PSU. For example, in a large SR PSU the segments are numbered in sample selection order from 1 to the largest number, say 12 in a smaller SR PSU. This approach uses a technique referred to as "combining strata" (see Kish, 1965, chapter 4) to group these segments into representative samples of the PSU. For example, to form two representative samples from 12 segments the segments with even numbers were collectively grouped to form SECU 1 for the SR PSU, and the remainder grouped to form SECU 2. Such pairs are "pseudo-strata" or sampling error strata formed for the purposes of variance estimation. Larger SR PSUs may be divided into more than two SECUs, and grouped to form more than one sampling error strata. For the remaining Non-Self-Representing (NSR) PSUs, an approach termed collapsing strata was used to form pairs of PSUs. There were 48 NSR PSUs available for the 2015-2017 NSFG. The strata from which the NSR PSUs were selected were inspected to identify pairs or strata that were as alike as possible, based on information on the sampling frame. The pairs were formed within Census divisions, and factors such as the socio-demographic and economic characteristics of the population in the strata were used in the pairing process.

In combination with the SR PSU SECUs, there are a total of 18 sampling error strata each with four SECUs for the 2015-2017 NSF. The number of strata and clusters is doubled for four-year datasets – that is, there are 36 sampling error strata with four SECUs in each stratum. A six-year dataset (2011-2013, 2013-2015, and 2015-2017) is also possible. Guidance on using the four-year and six-year weights for the various possible combinations of files can be found in the User's Guide for 2015-2017, Appendix 2, SAS and STATA Syntax Guidelines for Combining Data Across File Releases.

Table 12 presents estimates of standard errors for several key statistics from each of the two-year datasets. These estimates were developed using the variance estimation strategy described in this section.

Table 12. Estimated standard errors for four selected statistics, by race/ethnicity, age, and gender, 2011-2013, 2013-2015, and 2015-2017 NSFG.

	2011-2013			2013-2015			2015-2017		
Subgroup	*n*	Estimated percent	Standard error	*n*	Estimated percent	Standard error	*n*	Estimated percent	Standard error
Percent of current contraceptors who were using the oral contraceptive pill									
All	3,308	25.9	1.34	3,307	25.3	1.24	3,367	19.4	1.03
Hispanic	808	19.0	2.79	740	18.2	1.82	686	14.4	2.20
Non-Hispanic White	1,632	29.3	1.86	1,715	29.6	1.83	1,741	22.6	1.57
Non-Hispanic Black	702	18.2	2.15	644	20.2	2.43	769	13.8	1.94
Non-Hispanic Other	166	29.0	5.77	208	17.0	4.16	171	15.4	3.50
Percent of men who are married or cohabiting and intend to have a(nother) birth									
All, 15-49* years	1,215	58.1	1.90	1,191	58.7	1.94	1,172	56.1	1.98
15-19 years	19	88.8	8.97	10	79.1	12.3	7	90.1	10.92
20-24 years	150	84.2	3.13	133	81.6	3.47	105	88.3	4.25
25-29 years	323	79.0	3.71	306	82.0	2.69	252	85.4	3.16
30-34 years	311	61.5	3.59	329	64.5	3.68	274	65.5	4.10
35-39 years	226	46.7	3.95	235	40.4	5.00	238	56.7	5.01
40-44 years	186	17.6	3.51	178	21.0	3.76	159	22.6	4.52
45-49 years							137	8.2	2.46
Percent of females and males 15-19 who have ever had sexual intercourse									
Females 15-19 years of age	1,037	44.8	2.80	1,010	41.0	2.38	924	42.0	3.05
Males 15-19 years of age	1,088	47.1	2.22	999	42.1	1.93	886	37.8	2.95
Percent of single live births in the last 5 years that were breastfed at all									
All	1,657	75.3	1.97	2,387	77.4	1.93	1,267	81.9	1.79
Hispanic	498	76.4	3.61	692	86.0	2.38	320	90.8	2.22
Non-Hispanic White	711	78.4	2.60	1,021	75.5	2.90	558	82.2	1.99
Non-Hispanic Black	366	55.2	3.33	500	60.%	4.67	316	68.9	4.22
Non-Hispanic Other	82	86.1	5.19	174	90.4	2.32	73	71.9	8.33

***15-44 years for 2011-2013 and 2013-2015**

As in previous cycles of the NSFG, two variables, a sampling error stratum (SEST) and a sampling error computing unit (SECU), were added to the data set during routine processing for each case. The SEST and SECU codes were randomly assigned to eliminate virtually the chance of disclosure of geographic information that could adversely affect the privacy and confidentiality of respondent data. The SEST and SECU codes are mutually exclusive so that in a "stacked" dataset with four years of data they will be non-overlapping.

For a glossary of terms used in this document and related documents, see Appendix I in "2015-2017 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods."

11.References

American Association for Public Opinion Research (2015). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys, Eighth Edition*. Available at: http://www.aapor.org

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

Groves, R.M., and Heeringa, S. (2006). "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs." *Journal of the Royal Statistical Society, Series A*, 439-457.

Groves R.M., Mosher W.D., Lepkowski, J., and Kirgis, N.G. (2009). "Planning and Development of the Continuous National Survey of Family Growth." National Center for Health Statistics. *Vital Health Stat 1*(48).

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

Kish, L (1965). *Survey Sampling*. New York: John Wiley and Sons.

Kish, L. (1992). "Weighting for unequal P_i." *Journal of Official Statistics* **8**(2): 183-200.

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T.M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R.M., and Raghunathan, T.E. (2010). "Using Proxy Measures and Other Correlates of Survey Outcomes to Adjust For Non-Response: Examples From Multiple Surveys." *Journal of the Royal Statistical Society: Series A (Statistics in Society),* 173(2): 389-407.

Lepkowski, J. M., W. D. Mosher, R. M. Groves, B. T. West, J. Wagner and H. Gu (2013). Responsive Design, Weighting, and Variance Estimation in the 2006-2010 National Survey of Family Growth, National Center for Health Statistics. **2**.

Little, R. J. A. and S. Vartivarian (2005). "Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology* 31(2): 161-168.

Peytchev, A., Peytcheva, E., and Groves, R.M. (2010). "Measurement Error, Unit Nonresponse, and Self-Reports of Abortion Experiences." *Public Opinion Quarterly,* 74 **(**2): 319-327.

Wagner, J., B. T. West, H. Guyer, P. Burton, J. Kelley, M. P. Couper and W. D. Mosher (2017). The Effects of a Mid-Data Collection Change in Financial Incentives on Total Survey Error in the National Survey of Family Growth. Total Survey Error in Practice. P. P. Biemer, E. de Leeuw, S. Eckman, B. Edwards, F. Kreuter, L. E. Lyberg, N. C. Tucker and B. West, T. New York, Wiley.

Wagner, J., R. Valliant, F. Hubbard and L. Jiang (2014). "Level-of-Effort Paradata and Nonresponse Adjustment Models for a National Face-to-Face Survey." *Journal of Survey Statistics and Methodology* **2**(4): 410-432.

West, B. (2010). "An Examination of the Quality and Utility of Interviewer Estimates of Household Characteristics in the National Survey of Family Growth." Paper presented at the annual meeting of the American Association for Public Opinion Research, Chicago, May.

West, B. T., J. Wagner, F. Hubbard and H. Gu (2015). "The Utility of Alternative Commercial Data Sources for Survey Operations and Estimation: Evidence from the National Survey of Family Growth." *Journal of Survey Statistics and Methodology* **3**(2): 240-264.