2022-2023 National Survey of Family Growth: Study Design and Data Collection Procedures

Prepared Under Contract #75D30120C09732 for:
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20857

Prepared by:
Andy Peytchev, Jennifer Cooney, Taylor Lewis RTI International
3040 East Cornwallis Road
Research Triangle Park, NC 27709-2194

August 2025

Table of Contents

1. Introduction

This report provides key methodological information to users and analysts of the 2022-2023 National Survey of Family Growth (NSFG), beyond the information included in the User's Guide that accompanied the release of these public-use files in December 2024. The 2022-2023 NSFG includes 2 years (8 quarters) of data from the continuous NSFG. This 2-year period covers the first through eighth quarters of an overall 8-year (32-quarter) period of data collection from 2022 to 2029. It follows the prior 8-year period of continuous NSFG data collection from 2011 to 2019. See: 2011-2013 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods, 2013-2015 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods, 2015-2017 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods, and 2017-2019 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods for reports analogous to this one for the four data releases in that period.

This web-based report and related, detailed reports are intended to replace the Series 1 and Series 2 reports formerly published by the National Center for Health Statistics (NCHS) but still permit the timely release of essential information on the sample design and data collection methods for the NSFG.

From 1973 to 2002, NCHS conducted six periodic cycles of the NSFG before moving to a continuous survey design in 2006. This transition and new design have been described in prior reports:

- Planning and Development of the Continuous National Survey of Family Growth
- The 2006-2010 National Survey of Family Growth: Sample Design and Analysis of a Continuous Survey
- Responsive Design, Weighting, and Variance Estimation in the 2006-2010 National Survey of Family Growth

Changes were made for the 2011-2019 data collection period, described most recently in the 2017-2019 reports:

- 2017-2019 National Survey of Family Growth (NSFG): Summary of Design and Data Collection Methods
- 2017-2019 National Survey of Family Growth (NSFG): Sample Design Documentation
- 2017-2019 National Survey of Family Growth (NSFG): Weighting Design Documentation
- 2017-2019 National Survey of Family Growth (NSFG): Sample Error Estimation Design

The current report uses a similar format to the reports produced for the 2011-2019 survey to describe the new design for the 2022-2029 survey period, summarizes production outcomes for 2022-2023, and describes the weighting, variance estimation, imputation, and disclosure risk review and operations that took place to produce the public-use datasets.

2. Background on the National Survey of Family Growth

For background information on the purpose, content, and sponsorship of the NSFG, please see the main NSFG webpage, specifically the "About NSFG" section and the User's Guide for 2022-2023.

Sample design and data collection for the 2022-2023 NSFG were conducted by RTI International under a contract with NCHS, which covered 2020-2030. Data collection for the 2022-2023 survey began in January 2022 and continued through December 2023 yielding data files based on 2 years (or 8 quarters) of completed surveys. Interviews were conducted with a national probability sample of women and men 15-49 years of age living in households in the United States. The 9,957 completed interviews (5,586 with women and 4,371 with men) were administered by web, or in person by trained female interviewers using tablet computers, a procedure called computer-assisted personal interviewing (CAPI). A subset of the more sensitive questions was administered using computer-assisted self-interviewing (CASI). In this mode, respondents answered the questions on the tablet computer by themselves. In the web mode, respondents completed the entire survey, both CAPI and CASI sections, by themselves. In 2022-2023, the completed surveys for female respondents averaged 74 minutes in length, and those for male respondents average 48 minutes, both within the average survey lengths of 75 minutes for females and 50 minutes for males approved by the Office of Management and Budget (OMB No. 0920-0314). About 5% (452 of 9,957 completed surveys) were completed in Spanish, which is the only other language accommodated in the NSFG protocol.

3. Sample Design

This section of the document provides a brief overview of the 2022-2023 NSFG sample design. A more detailed description of the sampling procedures can be found in the **Sample Design** report.

The NSFG sampling procedures undertaken during the 2022-2023 NSFG were designed to meet several key objectives including the following:

1. minimizing the overall design effects for women and men,
2. controlling the costs of both screening and conducting the main survey,
3. targeting 10,500 survey completes overall,
4. oversampling teenagers aged 15-19 and non-Hispanic Black individuals of any age.

A nationally representative sample of the counties or groupings of counties, covering the entire land area of the 50 United States and the District of Columbia, was initially selected to cover the full 8-year data collection period of the continuous NSFG, and then randomly allocated into nationally representative annual subsamples. However, the analysis weights provided on data files are based on pooling 2 years of completed surveys.

3.1 Sample Universe

The target population, or population of inference, for the 2022-2023 NSFG consists of all non-institutionalized women and men aged 15-49 years as of the household screener, whose usual place of residence is the 50 United States or 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, long-term psychiatric hospitals, and those living on military bases. Included in the sample are

age-eligible college students living in group quarters (e.g., dormitories, sorority houses) and sampled through their parents' or guardians' households, and active-duty military personnel living off base.

3.2 Sample Selection

The NSFG is based on a stratified multistage area probability sample, using probability proportional to size (PPS) selection that oversamples areas with higher concentrations of age-eligible non-Hispanic Black individuals. There are five stages of sample selection:

1. Selection of primary sampling units (PSUs)
2. Selection of secondary sampling units (SSUs) within sampled PSUs
3. Selection of addresses within sampled SSUs
4. Selection of one eligible person within each sampled address
5. Two-phase sampling for nonresponse

These five stages are briefly outlined below. More details on sample selection are provided in the **Sample Design** report.

1. Selection of Primary Sampling Units

The **first stage** involved the selection of PSUs. PSUs were defined as counties or groupings of contiguous counties. The U.S. land area was divided into 2,023 PSUs on the sampling frame. The 22 Metropolitan Statistical Areas (MSAs) with at least 1 million (estimated) occupied housing units based on the 2015-2019 American Community Survey (ACS) were treated as certainty selections, meaning they were guaranteed to be included in one or more years of the 8-year data collection period. The remaining set of PSUs was constructed from individual counties or groupings of sparsely populated contiguous counties using a customized algorithm that seeks to minimize the spatial size of the PSU while maintaining sufficient sampling units therein.

The population of 2,023 – 22 = 2,001 non-certainty PSUs was stratified into $H = 8$ mutually exclusive groups based on the cross-classification of Census region and MSA status. In total, a sample of 222 unique certainty (n = 22) and non-certainty PSUs (n = 200) was selected and allocated across the 8-year data collection period. Some certainty PSUs appear in the sample in more than one year. For the 2022-2023 NSFG, 40 sampled PSUs were allocated to each data collection year: 25 non-certainty PSUs and 15 certainty PSUs. Non-certainty selections were made using a PPS sampling approach that oversampled areas of the county with higher concentrations of non-Hispanic Black individuals.

2. Selection of Secondary Sampling Units

In the **second stage** of selection, SSUs, also referred to as *segments*, were selected within PSUs. These were composed of Census Blocks Groups (CBGs) or groupings thereof, using the same general algorithm as for grouping PSUs. Like PSUs, SSUs were selected with PPS in a manner that oversamples SSUs with higher concentrations of non-Hispanic Black Individuals. With few exceptions, 12 SSUs were sampled within each PSU each year, and sets of 3 were allocated to each quarter.

3. Selection of Addresses Within SSUs

For the **third stage** of selection, addresses, a proxy for housing units, were selected within SSUs. The starting point in building the frame of housing units within each sampled SSU is RTI's enhanced address-based sampling (ABS) frame described at http://abs.rti.org/background. RTI's ABS frame is derived from the United States Postal Service's Computerized Delivery Sequence file and is updated monthly. Addresses are geocoded with latitude and longitude coordinates, effectively placing each address into a hierarchy of useful geographical designations, including CBGs.

Within each sampled SSU, the net coverage rate (Harter et al., 2021) of the ABS frame was estimated by way of the ratio of the count of locatable (i.e., city-style) addresses to the estimated count of occupied housing units using the most up-to-date ACS figures. For the 93% of SSUs with an estimated net coverage rate of at least 85%, addresses from the ABS frame were used as-is. For the remaining 7% of SSUs below that threshold, the sampling frame of addresses was populated by a dependent listing operation, whereby trained listers used electronic tablets loaded with an interactive, proprietary software containing street maps marked with the ABS-derived locatable housing units to confirm, remove, or add additional housing units to the SSU's frame.

For ABS (i.e., non-listed) SSUs, historical screening data from a prior survey was used to fit a logistic regression model to predict the probability that the given address contains one or more age-eligible individuals. Available covariates included those at an area level derived from the Census Planning Database and those at an address level that come from vendor data regularly appended to RTI's ABS frame. Address-level probabilities from the model were then ranked into three approximately equal-sized strata (low, mid, or high likelihood of age eligibility), and using a targeted eligibility rate boost of 4-5 percentage points, addresses in the higher likelihood strata were sampled at a higher rate relative to those in the lower likelihood strata.

4. Selection of Individuals within Addresses

For addresses containing one or more age-eligible individuals, a **fourth stage** of selection involved sampling one individual for the main survey. This was done via PPS, with a measure of size, and thus selection probability, for teens and females set to be larger than that of non-teens and males. These differential sampling probabilities were established to target 18.2% of all survey completes to be teen respondents and 55% of all survey completes to be female respondents.

5. Two-Phase Sampling for Nonresponse

As was done in prior NSFG survey years, the 2022-2023 NSFG used a two-phase sampling approach as a **fifth stage** of selection. Note that this refers to selection of a *sub*sample and other protocol changes from the first phase, as distinct from design phase as described below in Section 4. Each quarter, during the last 4 weeks of the 16-week data collection period, a subsample of nonresponding cases was selected for continued follow-up with an additional $40 promised incentive to participate. More details of this two-phase design are described in the **Sample Design** report.

4. Multimode, Multiphase, Continuous Data Collection Design

The 2022-2023 NSFG data collection marked the introduction of a web and face-to-face (FTF) multimode data collection design. The continuous data collection with four quarters per year was retained from the 2011-2019 NSFG. To accommodate a web-only data collection at the start of each quarter and to allow time for multiple mail invitations to the web survey, each quarterly data collection included three instead of two phases (as in 2011-2019) and continued for 16 weeks rather than 12 weeks, described in more detail below. Some of these key features of the 2022-2023 NSFG design are the following:

- Alignment to calendar years to facilitate data analysis and reporting.
- Three phases of data collection, with a change in protocol for each phase to increase participation.

- Initial phase with only web data collection to increase efficiency.
- Double sampling for nonresponse in the last phase for efficiency.
- Quarterly data collection with the ability to change sample sizes from one quarter to the next for flexibility.
- Overlapping quarters to maintain continuous work for field interviewers and supervisors during each year.

Screener and main survey data were collected solely by web during Phase 1, in the first 4 weeks of data collection. In-person data collection was introduced in Phase 2 in week 5 for 8 weeks, along with ongoing web data collection. Phase 3 started in week 13 for a subsample of screener and main survey nonrespondents, continuing to use web and in-person data collection, but with increased incentives. Screener and main survey data collection concluded at the end of week 16. These phases, durations, and schedule are shown in Figure 1.

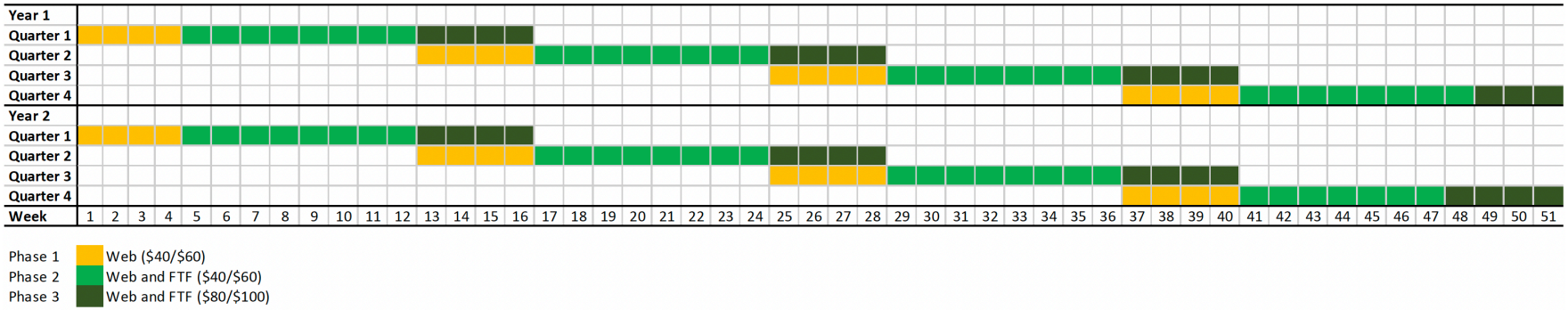

Figure 1. NSFG 2022-2023 Multimode Design

Phase 1: Web. In this phase, data were collected exclusively by web. A $2 token of appreciation and multiple mailings that varied in form and content were used in this phase, along with email and text invitations for the main survey.

Phase 2: FTF and Web. In-person data collection was introduced in this phase, while web data collection continued. A $40 incentive for the main survey was offered for both modes.

Phase 3: FTF and Web, Increased Incentive. A subsample of nonresponding households and individuals were selected for the more intensive data collection protocol in Phase 3. Compared to the 2017-2019 NSFG, the sampling rate was increased and a propensity model was not used to determine the selection probabilities, with both changes meant to reduce large Phase 3 weight variability and increase effective sample size. Sample addresses selected for this phase were sent a $5 screening incentive. Selected household members were offered an additional $40 for completion of the main survey, for a total of $80. An additional mailing was sent at the start of the phase to inform of these changes. Interviewers also had a reduced workload to focus on the selected nonresponding households.

Overlapping Quarters. The web-only data collection in Phase 1 was 4 weeks. Sequential quarters would have created an extended downtime for field interviewers and supervisors at the start of each quarter when data collection is only by web. It would have also allowed for only 8 weeks of field interviewing per quarter. The data collection design was created with overlapping quarterly samples to address these issues and allow sufficient time in data collection for the web-only mode to collect more data before implementing the more costly mode of data collection.

Field procedures were altered for the first quarter of data collection in 2022. The COVID-19 pandemic hindered in-person data collection for approximately 85% of the sample PSUs. There were also substantial related challenges in recruitment of in-person interviewers. As a result, the start of data collection was delayed by approximately 1 week with survey invitations mailed out on January 12, 2022. In-person data collection was started in Phase 3 instead of Phase 2.

5. Data Collection Activities

This section describes the protocols used for NSFG data collection. Interviewer training and equipment is briefly described, followed by the respondent contacting protocol, which included mailed and electronic invitation and reminder messaging for both modes of data collection, web and in-person. Key web and in-person survey data collection protocols are described, followed by the incentives utilized for both data collection modes. Lastly, experiments conducted during data collection are described. All survey instruments, operational plans, recruitment materials, and protocols were reviewed and approved by the Ethics Review Board (ERB) at NCHS, and OMB.

5.1 Interviewer Training

In 2022, the first year of data collection under the new contract, 63 field interviewers were trained and certified. In 2023, 78 new-to-project field interviewers and 21 veteran field interviews continuing from 2022 were trained and certified. All trainings and certifications were conducted virtually. Prior to training, field interviewers were sent a tablet computer preloaded with all required software and questionnaire instruments, hard copies of their field interviewer manual, showcard booklet, and other training materials.

New-to-project field interviewer training consisted of (1) self-study pretraining modules and (2) live instruction via teleconferencing software. Self-study pretraining modules included content on basic field

interviewing techniques and NSFG-specific content. Field interviewers also reviewed the NSFG Field Interviewer Manual and completed a quiz on its content. Live instruction occurred over 7 days and concluded with a required certification interview. Live instruction included content such as the following:

- effectively introducing the study on the doorstep and answering questions,
- completing the screener and main survey with selected eligible household members,
- administering consent and providing the token of appreciation to selected eligible household members, and
- explaining the use of Life History Calendar to selected eligible household members.

NCHS NSFG staff attended the live portion of new-to-project training. Bilingual field interviewers also attended an extra day of in-person instruction that was focused on administering the survey to Spanish-speaking respondents.

Veteran field interviewer training consisted of (1) self-study modules and (2) live instruction via teleconferencing software. The self-study modules include a refresher on key data collection protocols. Those protocols were reinforced during a virtual team meeting where field interviewers were also updated on key data collection protocol changes in 2023 and discussed how to address reluctance from selected households and eligible household members. Veteran field interviewer training concluded with an exercise where a group of FIs completed a scripted mock interview together.

5.2 Computer Hardware, Software, and Related Supplies

The household screeners and main surveys for 2022-2023 NSFG were conducted using computer-assisted web interviewing (CAWI) and CAPI. A CASI section was included in the CAPI (in-person) version of the main survey to give respondents an opportunity to answer a series of sensitive questions more privately (without the interviewer) using the tablet. All survey instruments were programmed in the Blaise software (version 5.12.8) developed by Statistics Netherlands.

The computers used for the 2022-2023 NSFG in-person data collection were Microsoft Surface Go 2 tablet computers. Field interviewer computer supplies included a detachable keyboard, protective case, AC adaptor, and power inverter. Interviewers were also provided with shredding scissors for secure disposal of the paper Life History Calendar.

5.3 Invitations and Reminder Contact Protocol

All selected housing units received the following mailings, prior to the initiation of in-person contact. Each mailing was sent approximately 4 business days apart.

1. An envelope containing an NSFG Advance Household Letter, an English NSFG Q&A Brochure, and a $2 bill. The letter was bilingual, with the content printed in English on one side and Spanish on the other.
2. A pressure-sealed self-mailer with content in both English and Spanish.
3. An envelope containing a reminder letter. The letter was bilingual, with the content printed in English on one side and Spanish on the other.
4. A folded reminder self-mailer postcard with content in both English and Spanish.
5. An envelope containing a reminder letter. The letter was bilingual, with the content printed in English on one side and Spanish on the other.

Throughout data collection, the following contact materials were sent to selected eligible household members:

- Invitation: Selected household members received an invitation letter that included instructions for accessing the web survey. The letter was bilingual, with the content printed in English on one side and Spanish on the other. If an email address was available, selected household members also received an email with similar content, and a text message if a phone number was available.
- Reminders: These were sent to selected respondents who had not started the main survey. A series of three reminder letters were sent, with each letter being dispatched 1 week apart. Once a respondent began the main survey, they no longer received any further reminder letters. Reminder letters were sent in batches twice weekly and were bilingual, with the content printed in English on one side and Spanish on the other. If an email address was available, selected household members also received an email with similar content and a text message if a phone number was available.
- Breakoff Reminders: These letters were sent to respondents who had started but not completed the main survey. Two breakoff letters with identical content were sent 1 week apart. Breakoff letters were sent in batches twice weekly and were bilingual, with the content printed in English on one side and Spanish on the other. If an email address was available, selected household members also received an email with similar content and a text message if a phone number was available.
- Parent Permission: These were sent to the parent or legal guardian of the selected minor respondent, to request their consent for the minor's participation. The parent permission letters were sent in batches once weekly and were bilingual, with the content printed in English on one side and Spanish on the other.

Throughout the field period of data collection, field supervisors requested that the following letters be mailed to select households:

- Unable to Contact: These letters were mailed when the field interviewer was not able to contact the household. If no one was reached after multiple attempts, a letter was sent notifying them of the contact attempts. If the household was in a gated community or apartment complex with restricted access, a letter was sent to the building authority explaining the purpose of the visit and introducing the interviewer. Unable to contact letters were sent in batches once weekly and had the content printed in English on one side and Spanish on the other.
- Screener Reluctance: These letters were mailed to households that were reluctant to participate in the screener. There were four versions of this letter, tailored to the type of reluctance encountered: general reluctance, not interested, too personal, and too busy. Screener reluctance letters were sent in batches once weekly and had the content printed in English on one side and Spanish on the other.
- Main Survey Reluctance: These letters were mailed to selected participants that were reluctant to participate in the main survey. There were five versions of this letter, tailored to the type of reluctance encountered: general reluctance, not interested, too personal, too busy, and parent/guardian reluctance (sent to the parent/guardian rather than the selected minor). Main reluctance letters were sent in batches once weekly and had the content printed in English on one side and Spanish on the other.

The following letters were mailed to nonrespondents who were subsampled for Phase 3 to encourage their participation:

1. Screener: These were sent to a subsample of screener survey nonrespondents. The envelope contained an advance household letter, an English NSFG Q&A Brochure, and a $5 bill. The letter was bilingual, with the content printed in English on one side and Spanish on the other.
2. Main Survey: These were sent to a subsample of main survey nonrespondents. The envelope contained a bilingual reminder letter, with the content printed in English on one side and Spanish on the other.

If an email and telephone number were available, selected individuals were also sent an email and text.

5.4 Web Surveying Protocol

The key steps in the protocol were the following:

1. As noted in Section 5.3, all sample households received multiple mailings that included instructions on completing the screener via the web, including a unique passcode assigned to the sample household. To complete the screener, a household member went to the study website, entered the passcode, and completed the brief (less than 5 minutes) screener. The screener then determined whether any household member was age-eligible for the survey. If more than one age-eligible household member was identified, the pre-programmed survey selection algorithm selected one person to complete the survey. If no one in the household was eligible, no further contact was made with the household.
2. Once selected to participate, **adult** respondents who were also the screener respondent were allowed to transition directly into the main survey. If the screener respondent was not the household member that was selected, contact information for the selected household member was collected. The selected household member then received an NSFG Advance Respondent Letter with instructions on how to complete the main survey on the web. This information was also sent via email within 24 hours if an email address was available for the respondent. Selected household members also received multiple prompts via letter and via email and text message if that information was available (see Section 5.3).
3. If a minor, defined as ages 15-17 in most states, was selected to participate in the main survey, the minor's parent or legal guardian was asked to provide and confirm their permission by typing their name and their teenager's name before the screener was considered complete. They were also offered the opportunity to print a copy of the parental permission form.
4. When an **adult** respondent logged into the main survey, they completed the consent process by reading the content and indicating whether they agreed to participate.
5. When a **minor** respondent, defined as ages 15-17 in most states, accessed the main survey, they were asked to read the assent form and to type in their name to register their assent to participate in the study.
6. After the main survey respondent provided their consent or assent to participate in the study, they were offered the option to receive their $40 token of appreciation via an emailed electronic gift card or mailed check.
7. While completing the main survey, respondents had access to multiple aids, including question-by-question guidance ("help screens") if additional information was needed on a particular question and the Life History Calendar (see page 35 of User's Guide), which was used only for female respondents as a tool to aid in recalling dates and detailed events.
8. Once the main survey was completed, respondents were sent their token of appreciation via their delivery method of choice (email or check).

5.5 In-person Protocol

The key steps in the protocol were the following:

1. Interviewers made visits to selected households predominantly during the evenings and weekends. If no one was home on the first visit, a "Sorry I Missed You" card and doorhanger was left at the household indicating that the interviewer had stopped by and would return at another time. Return visits were made to households during a different time of day or different day of the week than the initial contact.
2. When contact was made with a sample household, the field interviewer introduced herself to the household member by displaying her identification badge and identifying that she was an interviewer from RTI International contacting the household on behalf of the National Center for Health Statistics to discuss the National Survey of Family Growth. The NSFG Advance Household Letter was referenced and the NSFG Field Interviewer Letter of Authorization was shown if necessary.
3. If a household member was not willing or able to complete the screener at that time, the interviewer answered any questions regarding the study and the process and offered to return at a more convenient time.
4. After establishing that household member was an adult 18 or older and willing to participate in the brief (less than 5 minutes) screener, the interviewer conducted the household screener to determine whether any household member was age-eligible for the survey. If more than one age-eligible household member was identified, the pre-programmed survey selection algorithm selected one person to be interviewed. If no one in the household was eligible, no further contact was made with the household. Age was the primary basis for ineligibility; however, in some cases an age-eligible household member may have been ruled out based on their language or other factors. The NSFG screener and main survey could only be conducted in English or Spanish.
5. Once selected to participate, **adult** respondents were provided with an NSFG Advance Respondent Letter explaining that they had been selected for the survey and a copy of the NSFG Adult Consent Form covering all required elements of informed consent. An NSFG Q&A Brochure was also provided to address any frequently asked questions. The interviewer asked if they were willing to participate in the survey. If the respondent agreed, they were provided a $40 token of appreciation. The selected main survey respondent then provided an electronic signature acknowledging receipt of the token of appreciation.
6. For a **selected minor**, defined as ages 15-17 in most states, field interviewers obtained the permission of the minor's parent or legal guardian to talk with them. The parent or legal guardian was provided a copy of the NSFG Parent Permission Form and permission was confirmed by collecting their signature and teenager's name in the touchscreen tablet at the end of the screener. Once parent or legal guardian permission was obtained, the interviewer contacted the minor. The interviewer provided the minor with an NSFG Advance Respondent Letter and an NSFG Q&A Brochure, answered any questions the minor respondent had, and then asked if they were interested in participating. If the minor agreed, the interviewer would launch the survey instrument, provide the minor with a copy of the NSFG Minor Assent Form, ask the minor to provide an electronic signature and name acknowledging their assent, and then provide a $40 token of appreciation in advance of completing the interview.
7. The main survey was administered in a private setting with the interviewer reading the questions and entering the responses in the tablet. A private setting was defined as having no one over the age of 4 years within hearing range of the interviewer and respondent. Various aids were used throughout the interview: show cards that the respondent referred to for response categories; question-by-question guidance ("help screens") for the interviewer to read to the respondent if additional information was needed on a particular question; and the Life

History Calendar used only for female respondents as a tool to aid in recalling dates and detailed events.

8. The last section of the survey was completed solely by the respondent via CASI. The CASI section included questions on more sensitive topics. All respondents completed the CASI portion of the interview on their own and in such a way that the FI could not see the tablet screen. At the end of the CASI section, the respondent was prompted to lock the interview data before returning the computer to the interviewer. This locking made it impossible for the interviewer to back up and view any of the respondents' answers to CASI, nor could the interviewer back up and alter any prior responses to questions she administered before CASI. Questions included in the CASI section of the CAPI version of the main survey were also included in the CAWI version of the main survey, which was entirely self-administered.
9. Before leaving the household, the interviewer submitted the interview data, finalized the case, shredded the paper Life History Calendar with the project-provided shredding scissors (if applicable), and thanked the respondent for their participation.
10. During the respondent's CASI, the interviewer completed an NSFG Interview Observation Form. This form is used to collect information on how the overall interview went (where the interview took place, whether there were distractions, whether the female respondent used the Life History Calendar, etc.). After leaving the respondent's home and driving away, the interviewer found a safe, quiet place to enter the information from their paper interview observation form into the interview observation instrument on their tablet. The interviewer then shredded the paper interview observation form using the project-provided shredding scissors.

5.6 Use of Incentives

Households selected for Phase 1 were sent a $2 prepaid token of appreciation for completion of the screener. In Phase 1 and Phase 2, main survey respondents were offered a $40 token of appreciation, paid via an electronic gift card or check for web respondents and in cash for in-person respondents. Households selected for Phase 3 that were not yet screened in Phase 1 or 2 were sent a $5 prepaid token of appreciation for completion of the screener. In Phase 3, main survey respondents were offered an additional $40 (for a total of $80) as a token of appreciation for completion of the survey.

5.7 Experiments

One-page nonresponse follow-up (NRFU). One-page paper NRFU questionnaires were tested in quarters 1 and 2 of 2022, mailed to nonresponding households and nonresponding selected individuals. The paper NRFU was intended to assess how well it can be used to serve three objectives:

- identify ineligible households that can be removed from the denominator in response rate calculations,
- obtain measures of nonresponse bias for a select number of key variables, and
- consider using the select variables to inform another stage of main survey nonresponse weighting adjustments.

The paper NRFU was unsuccessful in accomplishing any of the three objectives because of very low return rate for the questionnaires. A possible explanation is that even though the request was for very little information compared to the full survey, mail mode is ineffective in the multimode design, particularly because households and individuals who have not responded by the end of Phase 3 have already been mailed survey invitations and reminders multiple times.

Paper screener. Offering a mail mode of data collection could help to increase response rates. Although a paper instrument for the NSFG main survey is not feasible, it is feasible for the NSFG screener. Half of the sample addresses in quarters 3 and 4 of 2022 were randomly assigned to receive a one-page paper screener in the third household mailing. The mail screener was not successful. Relatively few households returned the mail screener, and the screener or overall response rates were not increased. Combined with the additional cost of printing, processing, scanning, and other related labor, the mail screener was not implemented following the experiment.

Incentives. The $40 incentive for completion of the NSFG main survey in Phase 1 was first used in NSFG Cycle 6 (2002) and had remained the same for 2006-2010 and 2011-2019 data collection. Two higher incentive amounts were tested in quarters 3 and 4 of 2022 to compare to the $40 incentive: $60 and $80. Sample addresses were randomly assigned to one of the three incentive amounts. The additional $40 offered in Phase 3 of each quarter was retained. Response rates were higher for both $60 and $80 experimental conditions, but indistinguishable between each other. The experiment was continued in quarters 1-4 of 2023 with only two conditions, $40 and $60, assigning two-thirds of the sample to the $60 condition based on the results from 2022. Response rates were significantly higher for the $60 amount and improvement in sample balance and similarity to population distributions were seen for demographic variables for both the screener and main survey. Given this support for the $60 amount, it was approved for use starting in January 2024 (after data collection for this first 2-year period had ended).

Letter design. A household letter was designed based on visual design cues used in other fields and that have been demonstrated to elicit greater participation in surveys. The letter included the same content but was visually redesigned to show the respondent their progress in the process and their next steps with less reliance on whether they have read the full text in the letter. This letter was implemented for a random half of the sample in quarter 3 of 2023 and fully implemented in quarter 4 of 2023. Counter to prior studies, the visually redesigned letter did not increase participation.

QR code. A QR code was added to the advance household letter for the screener for a random half of the sample in quarter 4 of 2023. The purpose was to make it easier for the screener respondent to access the web screener. In addition to the NSFG survey URL, the QR code contained the passcode, simplifying the task to start the screener. The QR code increased screener completion particularly in Phase 1 (web), which was sustained through the end of the quarter.

6. Production Outcomes

The following series of tables show key production statistics from the 2022-2023 NSFG. Table 1 provides key summary counts for the overall NSFG sample and averages per quarter of data collection.

Table 1. Total number of sampled addresses, screened eligible households, and main NSFG surveys; and average number per quarter, 2022-2023 NSFG

	2022-2023
Sampled addresses[a]	
Total	98,307
Average per quarter	12,288
Screened eligible households[b]	
Total	17,818
Average per quarter	2,227
Main NSFG surveys[c]	
Total	9,957
Average per quarter	1,425

[a] 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 NSFG surveys refer to those completed by the selected age-eligible respondent from each sampled household. These numbers include partially completed surveys, which are those where the respondent at least reached the last applicable question before the final section, CASI for FTF respondents.

Table 2 shows selected indicators of fieldwork effort. The table shows the level of interviewer effort required to complete in-person main interviews in 2022-2023, at almost 17 hours per completed in-person main interview. Changes in these indicators will be tracked in future data releases under this new multimode design.

Table 2. Average number of calls (in-person visits) to obtain a screener, main survey, and the total, and average number of hours of interviewer labor to complete an in-person main survey, 2022-2023 NSFG

	Average number of calls
Number of screener calls to obtain in-person screener	1.90
Number of main survey calls to obtain in-person main survey[a]	2.0
Number of total calls to achieve in-person main survey[b]	3.90
Hours of Interviewer labor per completed in-person main survey	16.85

[a] Mean number of calls per in-person main survey is the average number of main calls on the cases with completed in-person main surveys.

[b] Mean number of total calls on a case to achieve in-person main survey is the average number of main and screener calls on the cases with completed in-person main surveys.

Table 3 shows the mean NSFG main survey length (for web and in-person surveys) overall, and by sex and age group. The mean survey length for females in 2022-2023 was 74 minutes. The mean survey length for males was 48 minutes. These survey lengths are similar to the 75 minutes for female surveys and 50 minutes for male surveys approved for NSFG by OMB.

Table 3. Mean and median length of main survey in minutes, for completed female and male surveys by age group: 2022-2023 NSFG

Sex and age	Mean[a] and median length of main survey in minutes	
	Mean	Median
Overall		
Total	62	56
Female		
Total	74	69
15-19	59	54
20-49	76	72
Male		
Total	47	44
15-19	44	40
20-49	48	44

[a] Excludes web survey item-level outliers that exceed 20 minutes on any question and web and in-person survey-level outliers that are less than 15 minutes or exceed 140 minutes for males and 180 minutes for females. Excludes partially completed surveys.

Table 4 contains the final case counts by sex, age, and race/Hispanic origin for the 2 years (8 quarters) of interview data included in the 2022-2023 NSFG. A total of 9,957 completed surveys or sufficient partial surveys were obtained in 2022-2023. Sufficient partials are respondents who at least answered the last applicable question before CASI for in-person respondents (or the final section of the survey for web respondents); some respondents may stop the survey then or stop somewhere during the CASI (final) section.

Table 4. Number of completed surveys in the 2022-2023 NSFG

	Number of completed surveys
Total	9,957
Sex	
Female	5,586
Male	4,371
Age	
15-19	1,448
20-49	8,509
Race/Hispanic origin	
Non-Hispanic Black	1,366
Hispanic	2,105
Non-Hispanic White and all other	6,486

[a] Counts include partial surveys, reaching a threshold deemed sufficiently complete to remain in the data file.

Finally, Table 5 shows the weighted response rates from the 2022-2023 NSFG, broken out by screener and main surveys and combined by multiplying the two together. The weighted screener response rate corresponds to the American Association for Public Opinion Research (AAPOR) Response Rate 3 (which is equivalent to Response Rate 4 since there are no partial screeners) and are weighted by the inverse of all selection probabilities through the address selection including the Phase 3 selection (AAPOR, 2023). Among the unscreened cases, the number that was eligible was estimated by using the eligibility rate from the screened cases within each phase and applied to the unscreened cases within each phase. The denominator of the screener response rate calculation was the number of known eligible cases in the sample plus the estimated number of eligible cases from unscreened households weighted by the

inverse of the selection probabilities including the Phase 3 selection. The weighted main response rate is conditional on screener completion and corresponds to the AAPOR Response Rate 4 where all cases are known eligible. All cases are weighted by the cumulative weights associated with the selection probabilities through the selection of an individual household member. The numerator was the number of main surveys (including partial main surveys), and the denominator was the number of completed screeners weighted by selection probability including the Phase 3 selection. The combined response rate is a simple multiplication of the screener response rate and the conditional main response rate. Table 6 shows the conditional weighted main survey response rates from the 2022-2023 NSFG, broken out by sex, race/Hispanic origin, and age.

Table 5. Weighted response rates by survey component: 2022-2023 NSFG

	Screener	Main	Combined
Total	49.4%	54.3%	26.8%

[a] All response rates have been weighted by the inverse of the probabilities of selection up through the specific stage of sampling relevant to each survey component including Phase 3 selection.

Table 6. Weighted main survey response rate conditional on screener response by sex, race/Hispanic origin, and age: 2022-2023 NSFG

	Female	Male
Total	54.7%	54.0%
Non-Hispanic Black	57.8%	55.5%
Hispanic	52.6%	49.3%
Non-Hispanic White and all Other races	55.0%	55.2%
Ages 15-19	48.0%	47.0%
Non-Hispanic Black	52.6%	50.9%
Hispanic	55.5%	50.5%
Non-Hispanic White and all Other races	44.8%	45.2%
Ages 20-49	55.8%	55.3%
Black	58.7%	56.3%
Hispanic	52.1%	49.1%
White and all Other races	56.6%	57.1%

7. Data Preparation for Public Use

For a description of the process by which the 2022-2023 survey data were cleaned, edited, and recoded in preparation for dissemination in public-use data files, see the User's Guide. Whereas the User's Guide provides a basic description of the imputation of recoded variables and the disclosure risk review process, the section below provides additional details on these preparations of the NSFG data for public use.

7.1 Imputation of Recodes

Most missing recode values were assigned using a multivariate hot deck imputation technique referred to as cyclical tree-based hot deck (CTBHD) (Sukasih & Scott, 2023). For each variable with missing data, an implicit tree-based model was fitted to create cells, or classes, that predict the values for the missing data based on other covariates in the data and logical consistency rules, where applicable. From within

these cells, a weighted sequential hot deck imputation procedure (Cox, 1980) is performed, which aims to preserve the weighted mean of the outcome within the cells before and after imputation. The procedure is termed "cyclical" because after the first fully imputed data set is created, it cycles back through the missing variables, in turn, and re-imputes missing values several times to build interdependence (Raghunathan et al., 2001) prior to terminating with a final imputed data set. The imputed values generated by CTBHD were checked to ensure that range constraints were maintained and that they were consistent with values populated for other related variables. Where necessary, manual edits were applied to achieve consistency.

In some cases, recodes were imputed using logical imputation, in contrast to the *model-based* CTBHD procedure just described. Logical imputation involved having a subject matter expert at NCHS examine variables related to the variable in question, and assign a value that was consistent with those other variables. In essence, logical imputation is an educated guess of the true value, which in some cases was akin to deductive imputation.

The recodes with the highest rate of imputation involved income. Hot deck imputation was used for just over 12% of cases for both poverty level (POVERTY) and total household income (TOTINCR). The imputation rate for other recodes did not exceed 3% of valid responses for the given variable.

Every NSFG recode variable that underwent imputation was assigned a corresponding imputation flag variable suffixed with "_I" indicating whether the value was based on questionnaire data, logical imputation, or model-based imputation. These flags allow users the flexibility to handle imputed cases as they may choose for their own analyses. However, it is the recommendation of NCHS that imputed values be retained in analyses to generate consistent point estimates for the population.

7.2 Procedures to Minimize Risk of Disclosure for Individual-level Data

Before any NSFG public-use file is released by NCHS, several disclosure risk reduction steps are taken to protect the confidentiality of respondents. First, NCHS staff provided specifications to RTI for modifying the data files for public use to prevent disclosure of the identities of the respondents. This included suppression of a significant number of century-month dates and other variables and the creation or collapsing of additional variables that could be used to identify small but visible groups or to match NSFG respondents with external data sets. In addition, the values of some observed variables were altered for an undisclosed portion of respondents in a process called statistical perturbation. Full details of this procedure are not disclosed to protect confidentiality and preserve the integrity of the process, but generally speaking, the process involved identifying candidate variables, deleting a small portion of substantive values reported, and re-imputing the missing values. The resulting values and distributions of each of the perturbed variables were then carefully checked to make sure that the recode specifications were satisfied. The process does not substantively alter any univariate point estimates, and effects on estimates of variance and tests of statistical significance are likewise minimal. The NSFG public-use files (PUFs) and restricted-use files (as made available in the Research Data Center) both contain the same perturbed values for these cases and variables. This perturbation step was accomplished concurrently with the model-based imputation of recodes described in Section 7.1.

Next, the proposed NSFG PUFs and related documentation were reviewed by the NCHS Disclosure Review Board (DRB), chaired by the NCHS Confidentiality Officer. Then, in response to the DRB's review, the NSFG staff and RTI made further changes where necessary to minimize the risk of disclosure. All of these changes are described in the User's Guide (see Section heading "Protections to Minimize Risk of Disclosure for Individual-Level Data") or in the PUF indices that show the full layout of the files for public use or the lists of restricted-use analytic variables to be made available only in the Research Data Center, all of which are posted to https://www.cdc.gov/nchs/nsfg/nsfg-2022-2023-puf.htm. In addition,

the codebook entries include a special note for each public-use variable where any disclosure risk reduction action has been taken.

7.3 Weighting and Variance Estimation

The development of weights and sample design indicators for variance estimation are briefly described here. For more detail, see reports on **Weighting Methodology** and **Variance Estimation**.

The final case weights for the 2022-2023 NSFG include (1) a base weight for the sampled address (i.e., housing unit) and individual selection probabilities, (2) a nonresponse adjustment to the address- and individual-level base weights, and (3) a calibration step to the latest population figures estimated from the 2022 ACS. The weight at the calibration step were trimmed to control the variance of the weights, since highly variable weights may inflate estimates of standard errors.

To adjust for any potential bias, nonresponse adjustment factors were developed. Sample-based unit nonresponse adjustments were developed by generating predicted probabilities of response using all available data for respondents and nonrespondents at the screener (i.e., address) and main survey (i.e., individual) levels. Separate models were fit to adjust for the screener and main survey, respectively, in part because there were slightly different data available at each level. Data on nonresponding cases to the screener included contextual data at the Census Block Group level and address-level indicators on the ABS frame. Data on nonrespondents to the main survey were derived from the screener instrument.

The last step of the weighting process is a calibration step that forces nonresponse-adjusted individual-level weights to match population totals. This can reduce sampling error and also may help reduce biases due to nonresponse or noncoverage. The selected variables/categories used for this step were age (in seven categories: 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, and 45-49), sex, and race/Hispanic origin (in four categories: White, Black, Hispanic, and Other), marital status (in three categories: married, previously married, and never married), and educational attainment (in two categories: high school or less, greater than high school).

The base probabilities of selection, nonresponse adjustments, and calibration adjustments were assimilated into a single, final weight: the variable WGT2022_2023. Extreme values of this weight were modestly trimmed to reduce the variability of the weights. We recommend that this weight variable be used for all analyses conducted from the 2-year file.

Table 7 shows the mean weights for key subgroups, along with the variance inflation caused by unequal weighting. After trimming, the minimum weight is 1,171 and the maximum is 95,540.

Table 7. Mean final weights (after post-stratification to Census data and trimming), and unequal weighting effect (UWE), by sex, age, and race/Hispanic origin, 2022-2023 NSFG

	Sample size	Mean weight	UWE
Total	9,957	15,129	1.70
Male	4,371	17,319	1.66
Female	5,586	13,415	1.69
15-19	1,429	14,976	1.45
20-49	8,528	15,154	1.74
Hispanic	2,105	15,720	1.74
Non-Hispanic Black	1,366	13,695	1.87
Non-Hispanic Other	6,486	15,239	1.65

In addition to differential weighting, the NSFG design is a stratified cluster sample. This stratification and clustering should be accounted for when estimating variance. To reflect the sample design as adequately as possible, without risking disclosure of the identity of respondents, we have created pseudo-strata and pseudo-clusters for variance estimation purposes. The clusters are identified by the variable VECL, and are numbered 1, 2, 3, and 4. These VECLs are nested within pseudo-strata identified by the variable VEST. That is, 80 unique VECLs are identified by the combinations of VEST and VECL codes, resulting in 60 complex survey degrees of freedom (Heeringa et al., 2017). These variables (VEST and VECL) should be used for any estimate of variance (see also Guidelines for Analysis below).

Table 8 shows estimated percentages and standard errors (reflecting the complex design) for four selected statistics, by race/Hispanic origin, age and sex, for 2022-2023 NSFG. These can be compared to estimates from 2002 NSFG and 2006-2010 NSFG Table X in Lepkowski et al. (2013) and estimates from 2011-2019 Table 9 in https://www.cdc.gov/nchs/data/nsfg/NSFG-2017-2019-Summary-Design-Data-Collection-508.pdf, but it is important to keep in mind that the prior survey estimates come from different survey designs and methodology that relied solely on in-person data collection. Also note that the figures presented in this table will not necessarily match those from similarly described recodes or other variables included in the public-use data files or used in published reports. This table will be updated following future PUF releases to compare counts and standard errors for similarly defined variables across this and the three additional future data releases planned for the 2022-2029 data collection period.

Table 8. Estimated percentages and standard errors for four key NSFG statistics, by selected characteristics 2022-2023 NSFG

Subgroup	*Unweighted n (in subgroup)*	Estimated (weighted) percent	Standard error
Percentage of female current contraceptors who were using the oral contraceptive pill, by race/Hispanic origin			
All female current contraceptors[1]	3,009	21.0	1.15
Hispanic	618	17.4	1.95
Non-Hispanic White	1,664	23.9	1.68
Non-Hispanic Black	385	14.0	2.24
Non-Hispanic Other	342	20.2	2.45
Percentage of men who intend to have a(nother) birth, by age			
All men ages 15-49	4,371	44.2	1.02
15-19	699	74.4	1.75
20-24	454	70.8	2.96
25-29	623	60.0	2.45
30-34	753	50.7	2.39
35-39	748	26.7	2.13
40-44	631	11.4	1.50
45-49	463	9.9	1.61
Percentage of female and male teenagers 15-19 who have ever had sexual intercourse			
Female teens	730	26.5	1.95
Male teens	699	29.8	1.99
Percentage of single live births in the last 5 years that were breastfed at all, by race/Hispanic origin of the mother			
All single births in last 5 years	1,618	81.9	2.02
Hispanic	340	84.4	3.38
Non-Hispanic White	848	84.9	2.09
Non-Hispanic Black	257	58.0	5.51
Non-Hispanic Other	173	91.7	2.82

[1]Contraceptors are defined as those who used any form of contraception in the month of the survey completion (CONSTAT1 codes 1-22).

8. Accounting for Complex Sample Design in Analysis: Examples of Program Statements

The data collected in the NSFG are obtained through a complex, multistage sample design that involves stratification, clustering, and oversampling of specific population subgroups. The final weights provided for analytic purposes have been adjusted in several ways to permit calculation of valid estimates for the noninstitutionalized, household-based population age 15-49 of the United States.

NSFG users are reminded that the use of standard statistical procedures based on the assumption that data are generated via simple random sampling (SRS) will generally produce *incorrect* estimates of variances and standard errors when used with NSFG data. Applying SRS techniques to NSFG data will generally produce standard error estimates that are, on average, too small, and are likely to generate results that are subject to excessive Type I error. For further details on analysis of complex sample survey data, see Lewis (2016), Heeringa et al. (2017), and Zimmer et al. (2025).

Analysts are strongly encouraged to use appropriate statistical software to reflect the complex sample design in their analyses. Several software packages are available for analyzing data collected from complex survey samples including SAS (https://www.sas.com/en_us/home.html), SUDAAN (https://www.rti.org/impact/sudaanr-statistical-software-analyzing-correlated-data), R (https://www.r-project.org/), and Stata (https://www.stata.com/). The key design variables for analysis are:

- VEST: Variance estimation stratum identifier
- VECL: Variance estimation cluster identifier
- WGT2022_2023: Analysis weight

Examples of program statements in SAS and Stata that illustrate the correct use of the design variables for variance estimation can be found on the webpage for the 2022-2023 public use release under the title "Variance Estimation Examples."

Below are additional examples of program statements for SUDAAN and R languages.

Example 1: Variance estimates for percentages using SAS-callable SUDAAN (11) and R (4.4.3) percentage of women ages 15-49 currently using the oral contraceptive pill, by age

Below are SAS-callable SUDAAN and R programs and output for an analysis of the percentage of women in the 2022-2023 NSFG female respondent file who were using the oral contraceptive pill during the month of interview. A cross-tabulation of use of the pill by age (15-19, 20-24, 25-29, 30-34, and 40-49) is generated.

The estimates and standard errors calculated are equivalent across SUDAAN and R and will also match those produced by SAS and Stata.

In these programs, variables in uppercase represent variables as named on the data files. Variables in lowercase represent variables that were created as part of this program. Library and file names are generic; the user must apply names specific to their computing environment.

SUDAAN 11

The DATA step creates a dataset for females that contains the variables to be used in the analysis, age categories (agerx), and current use of contraceptive pill (cpill). SUDAAN's PROC CROSSTAB produces a cross-tabulation of unweighted and weighted cell counts for the variables specified in the TABLES statement (agerx and cpill). Because these two variables are categorical, they must also appear in the

SUBGROUP statement and their respective numbers of categories specified in the LEVELS statement. The WEIGHT statement identifies the weight variable WGT2022_2023, and standard errors appropriate to the complex sample design are calculated by identifying the stratum identifier (VEST) first and the cluster identifier (VECL) second in the NEST statement. The DESIGN=WR option in the PROC statement invokes the ultimate cluster assumption that other software such as SAS uses by default to simplify variance estimation.

SUDAAN 11 Program

```
data EX1;
 set NSFG.females (keep=CASEID AGER CONSTAT1 VEST VECL WGT2022_2023);

if 15 le AGER le 19 then agerx=1;
 else if 20 le AGER le 24 then agerx=2;
 else if 25 le AGER le 29 then agerx=3;
 else if 30 le AGER le 34 then agerx=4;
 else if 35 le AGER le 39 then agerx=5;
 else if AGER ge 40 then agerx=6;

** Value of 6 on CONSTAT1 is oral contraceptive pill;
if CONSTAT1=6 then cpill=1;
 else cpill=2;
run;

* SUDAAN procedures require data set sorted by stratum and cluster codes;
proc sort data=EX1;
 by VEST VECL;
run;

PROC CROSSTAB DATA=EX1 FILETYPE=SAS DESIGN=WR;
NEST VEST VECL;
WEIGHT WGT2022_2023;
SUBGROUP agerx cpill;
LEVELS 6 2;
TABLES agerx*cpill;
PRINT NSUM WSUM ROWPER SEROW /
       WSUMFMT=F9.0 SEROWFMT=F6.3;
RUN;
```

SUDAAN 11 Output

```
                                   S U D A A N
 Software for the Statistical Analysis of Correlated Data
 Copyright Research Triangle Institute May 2020
 Release 11.0.4

DESIGN SUMMARY: Variances will be computed using the Taylor Linearization
Method, Assuming a With Replacement (WR) Design
 Sample Weight: WGT2022_2023
 Stratification Variables(s): VEST
 Primary Sampling Unit: VECL

Number of observations read : 5586 Weighted count : 74936918
Denominator degrees of freedom : 60

Date: 04-03-2025 SUDAAN Page: 1
Time: 09:10:51 Table: 1

Variance Estimation Method: Taylor Series (WR)
by: AGERX, CPILL.
```

		CPILL	
AGERX			
		Total	Yes	No
Total	Sample Size	5586	633	4953
	Weighted Size	74936918	8549165	66387753
	Row Percent	100.00	11.41	88.59
	SE Row Percent	0.000	0.643	0.643
15-19	Sample Size	730	96	634
	Weighted Size	10490061	1493243	8996818
	Row Percent	100.00	14.23	85.77
	SE Row Percent	0.000	1.679	1.679

```
-------------------------------------------------------------------------
| | | | | |
| 20-24 | Sample Size | 606 | 111 | 495 |
| | Weighted Size | 10927114 | 2071629 | 8855484 |
| | Row Percent | 100.00 | 18.96 | 81.04 |
| | SE Row Percent | 0.000 | 2.023 | 2.023 |
-------------------------------------------------------------------------
| | | | | |
| 25-29 | Sample Size | 803 | 116 | 687 |
| | Weighted Size | 10698176 | 1562548 | 9135628 |
| | Row Percent | 100.00 | 14.61 | 85.39 |
| | SE Row Percent | 0.000 | 1.889 | 1.889 |
-------------------------------------------------------------------------
| | | | | |
| 30-34 | Sample Size | 1028 | 113 | 915 |
| | Weighted Size | 11359278 | 1158344 | 10200933 |
| | Row Percent | 100.00 | 10.20 | 89.80 |
| | SE Row Percent | 0.000 | 1.384 | 1.384 |
-------------------------------------------------------------------------
| | | | | |
| 35-39 | Sample Size | 928 | 79 | 849 |
| | Weighted Size | 10977426 | 857487 | 10119939 |
| | Row Percent | 100.00 | 7.81 | 92.19 |
| | SE Row Percent | 0.000 | 1.105 | 1.105 |
-------------------------------------------------------------------------
| | | | | |
| 40-49 | Sample Size | 1491 | 118 | 1373 |
| | Weighted Size | 20484864 | 1405912 | 19078951 |
| | Row Percent | 100.00 | 6.86 | 93.14 |
| | SE Row Percent | 0.000 | 0.796 | 0.796 |
-------------------------------------------------------------------------
```

R 4.4.3

The read_sas() function in the "haven" package is used to read a SAS data set (.sas7bdat file) into the R session as a data frame named EX1. The mutate() function in the "dplyr" package is used to create age categories (variable agerx), while the base function ifelse() is used to create an indicator variable for current use of contraceptive pill (cpill).

The “survey” package of R features several functions prefixed by svy, which can be used to conduct complex survey data analyses properly accounting for stratification, clustering, and unequal weighting. The first step in utilizing them is to create a design object informing the package of these features. This is done below in the svydesign() function, which points to the stratum identifier (VEST), the cluster identifier (VECL), and the analysis weight (WGT2022_2023).

The svymean() and svytotal() functions are used to produce the marginal weighted proportions and totals (along with standard errors) for the two categorical analysis variables, and the svyby() function is used to produce the cross-tabulation of the two factors. Note how the “ex1_des” design object is always specified within these functions.

R 4.4.3 Program

```
### load necessary packages into R session
#install.packages('haven')
library(haven)

#install.packages('survey')
library(survey)

#install.packages('dplyr')
library(dplyr)

ex1 <- read_sas('/<path>/females.sas7bdat')

# create agerx variable, a grouping of integer AGER;
ex1$agerx <- ex1$AGER

ex1 <- ex1 %>%
 mutate(
 agerx = case_when(
 AGER >= 15 & AGER <= 19 ~ 1,
 AGER >= 20 & AGER <= 24 ~ 2,
 AGER >= 25 & AGER <= 29 ~ 3,
 AGER >= 30 & AGER <= 34 ~ 4,
 AGER >= 35 & AGER <= 39 ~ 5,
 AGER >= 40 ~ 6
 )
 )

# create cpill variable;
```

```
ex1$cpill <- ifelse(ex1$CONSTAT1 == 6, 1, 2)

# specify the complex survey design features
ex1_des <- svydesign(id=~VECL, strata=~VEST, weights=~WGT2022_2023,
 data=ex1,
 nest=T)

### marginal proportions and totals
# proportion of study population taking pill (1 = yes; 2 = no)
svymean(~as.factor(cpill), ex1_des, na.rm=FALSE)
# proportion of study population in each age grouping
svymean(~as.factor(agerx), ex1_des, na.rm=FALSE)

# estimated totals of study population taking pill (1 = yes; 2 = no)
svytotal(~as.factor(cpill), ex1_des, na.rm=FALSE)

# estimated totals of study population in each age grouping
svytotal(~as.factor(agerx), ex1_des, na.rm=FALSE)

### cross-classification of proportions and totals
# proportion of a particular age category taking pill
svyby(~as.factor(cpill), ~as.factor(agerx), ex1_des, svymean, se=T, na.rm=T)

# estimated totals of individuals in a particular age category taking pill
svyby(~as.factor(cpill), ~as.factor(agerx), ex1_des, svytotal, se=T, na.rm=T)
```

R 4.4.3 Output

```
> ### marginal proportions and totals
> # proportion of study population taking pill (1 = yes; 2 = no)
> svymean(~as.factor(cpill), ex1_des, na.rm=FALSE)
 mean SE
as.factor(cpill)1 0.11408 0.0064
as.factor(cpill)2 0.88592 0.0064
> # proportion of study population in each age grouping
> svymean(~as.factor(agerx), ex1_des, na.rm=FALSE)
 mean SE
as.factor(agerx)1 0.13999 0.0057
```

```
as.factor(agerx)2 0.14582 0.0063
as.factor(agerx)3 0.14276 0.0070
as.factor(agerx)4 0.15158 0.0063
as.factor(agerx)5 0.14649 0.0057
as.factor(agerx)6 0.27336 0.0077
>
> # estimated totals of study population taking pill (1 = yes; 2 = no)
> svytotal(~as.factor(cpill), ex1_des, na.rm=FALSE)
 total SE
as.factor(cpill)1 8549165 568019
as.factor(cpill)2 66387753 2651588
>
> # estimated totals of study population in each age grouping
> svytotal(~as.factor(agerx), ex1_des, na.rm=FALSE)
 total SE
as.factor(agerx)1 10490061 552116
as.factor(agerx)2 10927114 579466
as.factor(agerx)3 10698176 671258
as.factor(agerx)4 11359278 688812
as.factor(agerx)5 10977426 573133
as.factor(agerx)6 20484864 1065014
>
> ### cross-classification of proportions and totals
> # proportion of a particular age category taking pill
> svyby(~as.factor(cpill), ~as.factor(agerx), ex1_des, svymean, se=T,
na.rm=T)

      agerx       (cpill)1 (cpill)2 se.(cpill)1 se(cpill)2
1 1 0.14234840 0.8576516 0.016791804 0.016791804
2 2 0.18958614 0.8104139 0.020226476 0.020226476
3 3 0.14605740 0.8539426 0.018889272 0.018889272
4 4 0.10197342 0.8980266 0.013836002 0.013836002
5 5 0.07811371 0.9218863 0.011050002 0.011050002
6 6 0.06863176 0.9313682 0.007960796 0.007960796
>
> # estimated totals of individuals in a particular age category taking pill
> svyby(~as.factor(cpill), ~as.factor(agerx), ex1_des, svytotal, se=T,
na.rm=T)
```

```
      agerx       (cpill)1 (cpill)2 se.(cpill)1 se(cpill)2
1 1 1493243.4 8996818 207475.8 469101.3
2 2 2071629.3 8855484 232418.6 546690.7
3 3 1562547.8 9135628 217881.8 621944.8
4 4 1158344.4 10200933 170778.2 641460.7
5 5 857487.5 10119939 124594.6 554956.7
6 6 1405912.3 19078951 179174.0 1004109.3
```

Example 2: Variance estimates for percentages using SAS-callable SUDAAN (11) and R (4.4.3) mean number of children ever born, by urban/rural residence for women 15-49 years of age

Below are SAS-callable SUDAAN and R programs and output for an analysis of the mean number of children born to women 15-49 years of age in the 2022-2023 NSFG female respondent file, by urban/rural residence.

The estimates and standard errors calculated are equivalent across SUDAAN and R and will also match those produced by SAS and Stata.

In these programs, variables in uppercase represent variables as named on the data files. Variables in lowercase represent variables that were created as part of this program. Library and file names are generic; the user must apply names specific to their computing environment.

SUDAAN 11

The DATA step creates a dataset for females that contains the variables to be used in the analysis. SUDAAN's PROC DESCRIPT produces the weighted mean of the variable PARTIY overall and by the two levels of the classification variable urban (1 = urban; 2 = rural). The WEIGHT statement identifies the weight variable WGT2022_2023, and standard errors appropriate to the complex sample design are calculated by identifying the stratum identifier (VEST) first and the cluster identifier (VECL) second in the NEST statement. The DESIGN=WR option in the PROC statement invokes the ultimate cluster assumption that other software such as SAS uses by default to simplify variance estimation.

SUDAAN 11 Program

```
data EX2;
 set NSFG.FEMALES (keep=CASEID VEST VECL METRO PARITY WGT2022_2023);

if METRO in (1,2) then urban=1;
 else if METRO eq 3 then urban=2;
run;

* SUDAAN procedures require data set sorted by stratum and cluster codes;
proc sort data=EX2;
 by VEST VECL;
run;
```

```
PROC DESCRIPT DATA=EX2 FILETYPE=SAS DESIGN=WR;
NEST VEST VECL;
WEIGHT WGT2022_2023;
VAR PARITY;
SUBGROUP urban;
LEVELS 2;
PRINT NSUM MEAN SEMEAN LOWMEAN UPMEAN;
RUN;
```

SUDAAN 11 Output

```
 S U D A A N
 Software for the Statistical Analysis of Correlated Data
 Copyright Research Triangle Institute May 2020
 Release 11.0.4

DESIGN SUMMARY: Variances will be computed using the Taylor Linearization
Method, Assuming a With Replacement (WR) Design
 Sample Weight: WGT2022_2023
 Stratification Variables(s): VEST
 Primary Sampling Unit: VECL

Number of observations read : 5586 Weighted count : 74936918
Denominator degrees of freedom : 60

Date: 04-03-2025 SUDAAN Page: 1
Time: 09:46:05 Table: 1

Variance Estimation Method: Taylor Series (WR)
by: Variable, URBAN.

-------------------------------------------------------------------------
| | | URBAN |
| Variable | |-------------------------------------|
| | | Total | urban | rural |
-------------------------------------------------------------------------
| | | | | |
| Number of live | Sample Size | 5586 | 4796 | 790 |
```

```
| births | Mean | 1.11 | 1.06 | 1.38 |
| | SE Mean | 0.03 | 0.03 | 0.10 |
| | Lower 95% Limit | | | |
| | Mean | 1.05 | 0.99 | 1.19 |
| | Upper 95% Limit | | | |
| | Mean | 1.17 | 1.12 | 1.58 |
-------------------------------------------------------------------------------
```

R 4.4.3

The read_sas() function in the "haven" package is used to read a SAS data set (.sas7bdat file) into the R session as a data frame named EX2. The base function ifelse() is used to create an indicator variable for urban/rural residence.

The "survey" package of R features several functions prefixed by svy, which can be used to conduct complex survey data analyses properly accounting for stratification, clustering, and unequal weighting. The first step in utilizing them is to create a design object informing the package of these features. This is done below in the svydesign() function, which points to the stratum identifier (VEST), the cluster identifier (VECL), and the analysis weight (WGT2022_2023).

The svyby() function is used to estimate the weighted mean of the variable PARITY by the two classifications of residence (urban/rural). Note how it references the "ex2_des design" object. Ninety-five percent confidence intervals are computed using the general purpose confint() function, extracting the complex survey degrees of freedom (number of clusters – number of strata) from the same design object. This informs which reference t distribution to use in calculating endpoints.

R 4.4.3 Program

```
### load necessary packages into R session
#install.packages('haven')
library(haven)

#install.packages('survey')
library(survey)

ex2 <- read_sas('/<path>/females.sas7bdat')

# create urban/rural indicator;
ex2$urban <- ifelse(ex2$METRO == 1 | ex2$METRO == 2, "urban", "rural")

# specify the complex survey design features
ex2_des <- svydesign(id=~VECL, strata=~VEST, weights=~WGT2022_2023,
 data=ex2,
 nest=T)
```

```
# calculate mean of PARITY for the urban/rural indicator
means <- svyby(~PARITY, ~urban, ex2_des, svymean, se=T, na.rm=T)
means

# calculate confidence intervals for these means using complex survey df
CIs <- confint(means, df = degf(ex2_des))
CIs
```

R 4.4.3 Output

```
> # calculate mean of PARITY for the urban/rural indicator
> means <- svyby(~PARITY, ~urban, ex2_des, svymean, se=T, na.rm=T)
> means
 urban PARITY se
rural rural 1.384746 0.09826654
urban urban 1.056112 0.03318016
>
> # calculate confidence intervals for these means using complex survey df
> CIs <- confint(means, df = degf(ex2_des))
> CIs
 2.5 % 97.5 %
rural 1.1881832 1.581308
urban 0.9897418 1.122482
```

Example 3: Variance estimates for percentages using SAS-callable SUDAAN (11) and R (4.4.3) percentage of men 20-49 years of age who have ever had one or more biological children, by Hispanic origin and race

Below are SAS-callable SUDAAN and R programs and output for an analysis of the percentage of men aged 20-49 in the 2022-2023 NSFG male file who have ever fathered one or more biological children, tabulated by Hispanic origin and race.

The estimates and standard errors calculated are equivalent across SUDAAN and R, and will also match those produced by SAS and Stata.

In these programs, variables in uppercase represent variables as named on the data files. Variables in lowercase represent variables that were created as part of this program. Library and file names are generic; the user must apply names specific to their computing environment.

SUDAAN 11

The DATA step creates a dataset for males that contains a new variable indicating whether the respondent fathered one or more biological children (biokidsx) based on the variable EVBIOKID. For this

example, respondents who said “don’t know” or “refused” to answer EVBIOKID are coded as missing (sysmis) on biokidsx, but analysts may have different approaches. A subpopulation indicator for men ages 20-49 is also created. When producing estimates for population subgroups (such as men ages 20-49 as shown here), it is important not to subset the data, but instead create a subpopulation indicator variable (like the domain_flag variable used here) to identify the subgroup of interest within the given survey software. In SUDAAN, we specify this indicator variable (and applicable code for cases to retain) in the SUBPOPX statement.

SUDAAN’s PROC CROSSTAB produces a cross-tabulation of unweighted and weighted cell counts for the variables specified in the TABLES statement (HISPRACE2 biokidsx). As these two variables are categorical, they must also appear in the SUBGROUP statement and their respective numbers of categories specified in the LEVELS statement. The WEIGHT statement identifies the weight variable WGT2022_2023, and standard errors appropriate to the complex sample design are calculated by identifying the stratum identifier (VEST) first and the cluster identifier (VECL) second in the NEST statement. The DESIGN=WR option in the PROC statement invokes the ultimate cluster assumption that other software such as SAS uses by default to simplify variance estimation.

SUDAAN 11 Program

```
data EX3;
 set NSFG.MALES (keep=CASEID EVBIOKID AGER HISPRACE2
                  VEST VECL WGT2022_2023);

biokidsx=0;
if EVBIOKID eq 1 then biokidsx=1;
 else if EVBIOKID in (8 9) then biokidsx=.;

**create a variable for subpopulation of ages 20 and older;
agepop=2;
if AGER ge 20 then agepop=1;

if agepop=1 and biokidsx in(1 2) then domain_flag=1;
 else domain_flag=0;
run;

* SUDAAN procedures require data set sorted by stratum and cluster codes;
proc sort data=EX3;
 by VEST VECL;
run;

PROC CROSSTAB DATA=EX3 FILETYPE=SAS DESIGN=WR;
NEST VEST VECL;
WEIGHT WGT2022_2023;
```

```
SUBPOPX domain_flag=1;
SUBGROUP HISPRACE2 biokidsx;
LEVELS 4 2;
TABLES HISPRACE2*biokidsx;
PRINT NSUM WSUM ROWPER SEROW /
       WSUMFMT=F9.0 SEROWFMT=F6.3;
RUN;
```

SUDAAN 11 Output

```
 S U D A A N
 Software for the Statistical Analysis of Correlated Data
 Copyright Research Triangle Institute May 2020
 Release 11.0.4

DESIGN SUMMARY: Variances will be computed using the Taylor Linearization
Method, Assuming a With Replacement (WR) Design
 Sample Weight: WGT2022_2023
 Stratification Variables(s): VEST
 Primary Sampling Unit: VECL

Number of observations read : 4371 Weighted count : 75700206
Observations in subpopulation : 3644 Weighted count : 64371210
Denominator degrees of freedom : 60

Date: 04-03-2025 SUDAAN Page: 1
Time: 09:56:54 Table: 1

Variance Estimation Method: Taylor Series (WR)
For Subpopulation: DOMAIN_FLAG = 1
by: Race and Hispanic origin - based on 1997 OMB guidelines, BIOKIDSX.

-------------------------------------------------------------------------
| | | BIOKIDSX |
| Race and | |----------------------------------|
| Hispanic origin | | Total | one or | none |
| - based on 1997 | | | more | |
| OMB guidelines | | | | |
```

```
-------------------------------------------------------------------------
| | | | | |
| Total | Sample Size | 3644 | 1448 | 2196 |
| | Weighted Size | 64371210 | 29718526 | 34652683 |
| | Row Percent | 100.00 | 46.17 | 53.83 |
| | SE Row Percent | 0.000 | 1.232 | 1.232 |
-------------------------------------------------------------------------
| | | | | |
| 1) HISPANIC | Sample Size | 672 | 266 | 406 |
| | Weighted Size | 13979964 | 6789912 | 7190052 |
| | Row Percent | 100.00 | 48.57 | 51.43 |
| | SE Row Percent | 0.000 | 2.510 | 2.510 |
-------------------------------------------------------------------------
| | | | | |
| 2) NON-HISP WH | Sample Size | 2026 | 797 | 1229 |
| | Weighted Size | 35075287 | 15776165 | 19299122 |
| | Row Percent | 100.00 | 44.98 | 55.02 |
| | SE Row Percent | 0.000 | 1.722 | 1.722 |
-------------------------------------------------------------------------
| | | | | |
| 3) NON-HISP BL | Sample Size | 433 | 188 | 245 |
| | Weighted Size | 7502225 | 3774383 | 3727842 |
| | Row Percent | 100.00 | 50.31 | 49.69 |
| | SE Row Percent | 0.000 | 3.016 | 3.016 |
-------------------------------------------------------------------------
| | | | | |
| 4) NON-HISP OTH | Sample Size | 513 | 197 | 316 |
| | Weighted Size | 7813733 | 3378066 | 4435667 |
| | Row Percent | 100.00 | 43.23 | 56.77 |
| | SE Row Percent | 0.000 | 3.373 | 3.373 |
-------------------------------------------------------------------------
```

R 4.4.3

The read_sas() function in the "haven"' package is used to read a SAS data set (.sas7bdat file) into the R session as a data frame named EX3. A new variable indicating whether the respondent fathered one or more biological children is created (biokidsx) based on the variable EVBIOKID. For this example, respondents who said "don't know" or "refused" to answer EVBIOKID are coded as missing (i.e., NA) on

biokidsx, but analysts may have different approaches. A subpopulation indicator for men ages 20-49 is also created.

The "survey" package of R features several functions prefixed by svy, which can be used to conduct complex survey data analyses properly accounting for stratification, clustering, and unequal weighting. The first step in utilizing them is to create a design object informing the package of these features. This is done here in the svydesign() function, which points to the stratum identifier (VEST), the cluster identifier (VECL), and the analysis weight (WGT2022_2023). When producing estimates for population subgroups (such as men ages 20-49 as shown here), it is important not to subset the data, but instead create a subpopulation indicator variable (like the domain_flag variable used here) and either use to subset the design object or use as part of the svyby() function. Both techniques are illustrated below.

R 4.4.3 Program

```
### load necessary packages into R session
#install.packages('haven')
library(haven)

#install.packages('survey')
library(survey)

ex3 <- read_sas('/<path>/males.sas7bdat')

# create biokidsx variable
ex3$biokidsx <- 0
ex3$biokidsx[ex3$EVBIOKID == 1] <- 1
ex3$biokidsx[ex3$EVBIOKID %in% c(8, 9)] <- NA

# create domain indicator variable;
ex3$domain_flag <- ifelse(ex3$AGER >= 20 & (ex3$biokidsx == 1 | ex3$biokidsx
== 0), 1, 0)

# specify the complex survey design features
ex3_des <- svydesign(id=~VECL, strata=~VEST, weights=~WGT2022_2023,
 data=ex3,
 nest=T)

# create a subset of this design
ex3_des_sub <- subset(ex3_des, ex3$domain_flag == 1)

### get marginal proportions and totals
# HISPRACE2 proportions for domain
```

```
svymean(~as.factor(HISPRACE2), ex3_des_sub, na.rm=FALSE)
# weighted totals of HISPRACE2 for domain
svytotal(~as.factor(HISPRACE2), ex3_des_sub, na.rm=FALSE)
# estimated proportions of biokidsx for domain
svymean(~as.factor(biokidsx), ex3_des_sub, na.rm=FALSE)
# weighted totals of biokidsx for domain
svytotal(~as.factor(biokidsx), ex3_des_sub, na.rm=FALSE)

### get cross-classification of proportions and totals
# proportions of biokidsx by HISPRACE2 categories for domain
svyby(~as.factor(biokidsx), ~as.factor(HISPRACE2), ex3_des_sub, svymean,
se=T, na.rm=T)
# weighted totals of biokidsx by HISPRACE2 categories for domain
svyby(~as.factor(biokidsx), ~as.factor(HISPRACE2), ex3_des_sub, svytotal,
se=T, na.rm=T)
```

R 4.4.3 Output

```
> ### get marginal proportions and totals
> # HISPRACE2 proportions for domain
> svymean(~as.factor(HISPRACE2), ex3_des_sub, na.rm=FALSE)
 mean SE
as.factor(HISPRACE2)1 0.21718 0.0224
as.factor(HISPRACE2)2 0.54489 0.0224
as.factor(HISPRACE2)3 0.11655 0.0099
as.factor(HISPRACE2)4 0.12139 0.0147
> # weighted totals of HISPRACE2 for domain
> svytotal(~as.factor(HISPRACE2), ex3_des_sub, na.rm=FALSE)
 total SE
as.factor(HISPRACE2)1 13979964 1589950
as.factor(HISPRACE2)2 35075287 1936390
as.factor(HISPRACE2)3 7502225 650372
as.factor(HISPRACE2)4 7813733 1116169
> # estimated proportions of biokidsx for domain
> svymean(~as.factor(biokidsx), ex3_des_sub, na.rm=FALSE)
 mean SE
as.factor(biokidsx)0 0.53833 0.0123
as.factor(biokidsx)1 0.46167 0.0123
> # weighted totals of biokidsx for domain
```

```
> svytotal(~as.factor(biokidsx), ex3_des_sub, na.rm=FALSE)
 total SE
as.factor(biokidsx)0 34652684 1755313
as.factor(biokidsx)1 29718526 1521689
>
> ### get cross-classification of proportions and totals
> # proportions of biokidsx by HISPRACE2 categories for domain
> svyby(~as.factor(biokidsx), ~as.factor(HISPRACE2), ex3_des_sub, svymean,
se=T, na.rm=T)
 HISPRACE2 (biokidsx)0 (biokidsx)1 se(biokidsx)0 se(biokidsx)1
1 1 0.5143112 0.4856888 0.02509542 0.02509542
2 2 0.5502199 0.4497801 0.01721714 0.01721714
3 3 0.4968982 0.5031018 0.03016136 0.03016136
4 4 0.5676758 0.4323242 0.03372886 0.03372886
> # weighted totals of biokidsx by HISPRACE2 categories for domain
> svyby(~as.factor(biokidsx), ~as.factor(HISPRACE2), ex3_des_sub, svytotal,
se=T, na.rm=T)
 HISPRACE2 ( biokidsx)0 (biokidsx)1 se(biokidsx)0 se(biokidsx)1
1 1 7190052 6789912 823442.7 909278.9
2 2 19299122 15776165 1203242.0 1079723.5
3 3 3727842 3774383 391395.7 400927.8
4 4 4435667 3378066 674037.4 561211.1
```

9. References

American Association for Public Opinion Research. (2023). *Standard definitions: Final dispositions of case codes and outcome rates for surveys. 10th Ed*. AAPOR. Available at www.aapor.org.

Cox, B. G. (1980). The weighted sequential hot deck imputation procedure. *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 721–726.

Harter, R., Morton, K., Amaya, A., & Brown, D. (2021). Estimating net coverage of segments in ABS frames. *Field Methods, 33*, 68-84.

Heeringa, S. G., West, B. T., & Berglund, P. A. (2017). *Applied survey data analysis*. 2nd Ed. Boca Raton, FL: CRC Press.

Lepkowski, J. M., Mosher, W. D., Groves, R. M., West, B. T., Wagner, J., & Gu, H. (2013). *Responsive design, weighting, and variance estimation in the 2006-2010 National Survey of Family Growth*. Vital and Health Statistics, Series 2, No. 158. Hyattsville, MD: National Center for Health Statistics. Available at http://www.cdc.gov/nchs/data/series/sr_02/sr02_158.pdf.

Lewis, T. (2016). *Complex survey data analysis with SAS®*. Boca Raton, FL: Chapman and Hall/CRC.

Raghunathan, T., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, *27*(1), 85-95.

Sukasih, A. S., & Scott, V. (2023). *Cyclical tree-based hot deck imputation*. RTI Press. RTI Press Methods Report No. MR-0052-2307 https://doi.org/10.3768/rtipress.2023.mr.0052.2307

Zimmer, S., Powell, R., & Velásquez, I. (2025). *Exploring complex survey data analysis using R: A tidy introduction with {srvyr} and {survey}*. CRC Press. https://tidy-survey-r.github.io/tidy-survey-book/

10. Appendix 1: Glossary

ABS frame, or Address-Based Sampling frame—The RTI address-based sampling frame is derived from the Delivery Sequence File from the U.S. Postal Service, which lists all addresses to which mail is currently delivered by the Postal Service. The ABS frame was periodically updated. The ABS frame was used to select the 2022-2023 NSFG sample addresses. In some areas, the ABS sample was augmented with listing.

Blaise—A software system developed by Statistics Netherlands, which was used to program the NSFG questionnaires for use with in-person and web respondents. The Blaise survey instruments route the respondent to the next appropriate question, store the respondent's answers, and permit checking the consistency of one answer with answers to other related questions. Blaise has been used for the NSFG since 1995. The 2022-2023 NSFG was the first NSFG to use Blaise for both in-person and web data collection.

Call—In-person visit by an interviewer to a housing unit in the NSFG sample. Household calling for screener and main interviews was done only in person in the NSFG. Some calls resulted in a *contact* (speaking with someone in the household), while other calls resulted in no contact (either the address was not occupied or no one was at home). Thus, calls represent any visit, regardless of outcome.

CAPI—Computer-assisted personal interviewing, in which the interviewer used a laptop computer to administer questions in the interview. The laptop displayed question text for the interviewer to read and provided any other necessary instructions to the interviewer. Interviewers recorded the respondent's answers using the keyboard. Software directed the interviewer to the next appropriate question based on the answers entered.

CASI—Computer-assisted self-interviewing. For in-person respondents in the NSFG, a portion of the main survey was completed by the respondent using CASI after all the interviewer-administered sections. The respondent chose a desired response option to each question, using the laptop keyboard. The software directed the respondent to the next appropriate question based on the answers entered. As in all past NSFGs that were computerized, the in-person respondent in the 2022-2023 NSFG performed these steps privately while the interviewer completed other tasks nearby, to offer the respondent as much privacy as possible. (Note: Web respondents, for whom the entire main survey was self-administered, answered largely the same questions in the final section of their survey.)

DRB, or Disclosure Review Board—A committee of peer reviewers evaluating all proposed public-use files for the potential for unintended identification of individuals linked to their data and assessing procedures to protect against such disclosure.

DSF, or Delivery Sequence File—The Delivery Sequence File from the U.S. Postal Service lists all addresses to which mail is currently delivered by the Postal Service. The DSF is the basis of the ABS frame. See "ABS."

Double (or two-phase) sample—A subsample of nonresponding sample cases (either at the screener stage or the main interview stage), selected for further follow-up efforts after the completion of the first phase of data collection. The design for 2022-2023 subsampling was implemented for design phase 3. NSFG has used such a subsample follow-up approach since the 2002 survey.

Electronic life history calendar—For the 2022-2023 NSFG an electronic version of the life history calendar was included for the web mode. See "Life history calendar."

Eligible household—A household containing at least one person who is eligible for the NSFG—that is, males or females 15-49 years of age at the date on which the screener was completed, and living in the household population of the United States (all 50 states or the District of Columbia). It is not known whether a selected household has an eligible person until the household screener is conducted. If a household had two or more persons 15-49 years of age, one of these persons was selected randomly for the NSFG main interview.

Eligibility rate—The percentage of sample cases that were members of the target population. In the 2022-2023 NSFG the eligibility rate was the percentage of households that contained a person aged 15-49.

ERB, or Ethics Review Board—A committee of peer reviewers of research procedures involving human subjects that weighs the benefits of the research relative to the risks of harm to human subjects.

Item imputation—The process of assigning data values to cases with missing data ("don't know," "refused," or "not ascertained") for a particular variable. In the NSFG, item imputation was done for a small subset of "recoded variables," or "recodes" (defined below, under "recodes"), rather than all of the thousands of variables in the data set. The purpose of imputation is to make the data more complete, more consistent, and easier to use, and, most importantly, to reduce bias caused by differential nonresponse. For example, if a respondent did not report their educational level, the education level would get imputed, perhaps to a value of "high school graduate." Imputation was done in two ways in the 2022-2023 NSFG, logical and model-based weighted sequential hot deck imputation. Hot deck imputation assigns a value for a case with missing data from a similar respondent with a non-missing value. Hot deck imputation was used to assign most of the imputed values. Occasionally, however, logical imputation was used. Logical imputation uses a subject matter expert to assign a value based on the value of other variables for the case with missing data.

Life history calendar—A visual presentation of a calendar covering the reference period of various questions in the female survey for NSFG, used to help the respondent record key personal events used as landmark events to cue memories of the dates of events measured in the survey. In the 2022-2023 NSFG the female in-person interview included a paper life history calendar for respondents to use as a recall aid for sections of the interview with more challenging recall tasks, such as the pregnancy and contraceptive history sections. An electronic life history calendar was used for female respondents in the web mode.

Main interview (or main survey)—An interview sought with or a survey completed by the selected household member within sample households containing an eligible target population member. If the household screener revealed that the household contained one or more persons 15-49 years of age, a main interview/survey was requested from one of those persons. If there were two or more persons 15-49, one such person was selected at random for the main survey, based on pre-programmed selection probabilities.

Measure of size—A value assigned to every sampling unit in a sample selection. Typically measures of size are a count of units associated with the elements to be selected. This allows different probabilities of selection across the various units of unequal sizes. For a description of the measures of size used by the 2022-2023 NSFG, please see the **Sample Design** report.

Multiphase design—A survey design that changes its sample design or recruitment protocol over different sets of sample cases or over time periods of the survey, to obtain optimal balance of costs and quality of survey estimates.

Multistage sample design—A sample design that selects units at different levels and points in the data collection process, such as PSUs, SSUs, households, and individuals.

National Center for Health Statistics—NCHS is the United States' principal health statistics agency. It designs, develops, and maintains a number of systems that produce data related to demographic and health concerns. These include data on registered births and deaths collected through the National Vital Statistics System, the National Health Interview Survey, the National Health and Nutrition Examination Survey, the National Health Care Surveys, and the NSFG, among others. NCHS has conducted the NSFG since 1973. NCHS is one of the "Centers" for Disease Control and Prevention (CDC), which is part of the U.S. Department of Health and Human Services.

Office of Management and Budget Clearance—OMB reviews survey materials and questionnaires proposed for use by government agencies under the provisions of the Paperwork Reduction Act. The review is conducted by OMB's Office of Information and Regulatory Affairs. No survey of more than nine persons can be conducted by a U.S. government agency without review, including federal register notices for public comment, and approval by OMB.

Phase—A period of data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols, and measurement conditions are used. Starting with the 2022-2023 NSFG, the NSFG used a three-phase approach combined with 16-week quarters. Phase 1 consisted of weeks 1-4, in which only web data collection was used. Phase 2 consisted of weeks 5-12, in which in-person data collection was added to the web mode. Phase 3 consisted of weeks 13-16, in which a subsample of nonrespondents from phase 2 was offered higher incentives and field effort was focused on the selected subsample for either in-person or web survey completion.

Public-use file—An electronic data set containing records based on information collected from survey respondents. These files typically include a subset of variables collected in the survey that have been reviewed extensively (by survey staff and NCHS DRB) to ensure that the identities of the respondents are protected. These files are disseminated by NCHS to fulfill its obligations to provide data files for public use as part of the federal statistical system and to fulfill its obligations to the cosponsoring agencies.

PSU—Primary sampling unit. The first stage selection unit in a multistage area probability sample. In the NSFG, PSUs are counties or groups of counties in the United States; 222 PSUs were selected into the NSFG sample for 2022-2029. There were 40 PSUs in the 2022-2023 sample.

Recodes or recoded variables—It is not possible to edit or impute all of the variables in the continuous NSFG data file. NSFG staff selected a subset of variables from the NSFG data file that were constructed, edited, and imputed. These are called recodes or recoded variables. Recodes are variables that are likely to be used frequently by NCHS and other data users. They were edited for consistency, and missing values were imputed. Many (but not all) of these recoded variables were constructed from other variables in the NSFG; some were constructed from a large number of other variables. Other variables in the data file were not edited or imputed in this way.

Respondent—A person selected into the sample who provides a complete or partial survey. In the 2022-2023 NSFG, the "respondents" are the 5,586 women and 4,371 men 15-49 years of age who completed the NSFG main survey.

Response rate—Respondents to a survey divided by the number of eligible persons in the sample. In this report, the response rate is the number of respondents (15-49 years of age) divided by the number of eligible persons (15-49 years of age) in the sample. Given that not all screeners were completed, the number of eligible persons is not known precisely, so this number is estimated.

RTI—RTI International is a nonprofit research organization that developed the multimode data collection design, conducted the web administration and fieldwork, and data processing for the 2022-2023 NSFG under a contract with NCHS.

Sampling variance—The sampling variance is a measure of the variation of a statistic, such as a proportion or a mean, which is the result of having selected a random sample instead of collecting data from every person in the full population. It measures the variation of the estimated proportion or mean over repeated samples. The sampling variance is zero when the full population is observed, as in a census. For the NSFG, the sampling variance estimate is a function of the sampling design and the population parameter being estimated (for example, a proportion or a mean). Many common statistical software packages compute "population" variances by default; these may underestimate the sampling variance. Estimating the sampling variance requires special software, such as those discussed in this report.

Sampling weight—For a respondent in the NSFG, the estimated number of persons in the target population that they represent. For example, if a man in the sample represents 12,000 men in his age and race/Hispanic origin category, then his "sampling weight" is 12,000. The NSFG sampling weights adjust for different sampling rates (of the age and race/Hispanic origin groups), different response rates, and different coverage rates among persons in the sample, so that accurate national estimates can be made from the sample. Because it adjusts for all these factors, it is sometimes called a "fully adjusted" sampling weight.

Screening survey—Sometimes called a "household screener," a screening survey is a (usually short) set of questions, asked of a household informant with the chief goal of determining whether the household contains anyone eligible for the survey. In the NSFG, the screening survey consisted of a short household roster, collecting age, race, Hispanic origin, and sex. Those households having one or more persons 15-49 years of age were eligible for a main interview. In the NSFG, only persons 18 and older could be screener informants.

Self-representing area—A county or group of counties forming a primary sampling unit with population counts sufficiently large to be equal to or greater than the typical stratum sample size in the U.S. national sample. Such PSUs were thus represented in all draws of a national sample using the design. The sampling probabilities for persons in such areas were designed to be equal to those applicable in smaller PSUs, called non–self-representing areas.

SRS, or Simple random sample—A sample in which all members of the population are selected directly and have an equal chance to be selected for the sample. The NSFG sample was not a simple random sample. The NSFG sample was stratified, selected in stages, and employed unequal chances of selection for the respondents, varied by age, race/Hispanic origin, and sex. Such designs are referred to as "complex" and require special software to estimate the variance of statistics computed from a sample with a complex design.

SSU—Secondary sampling unit. A group of housing units located near one another, all of which were selected into the sample. Also referred to as "segment."

Strata; stratification—Stratification is the partitioning of a population of sampling units into mutually exclusive categories (strata). Typically, stratification is used to increase the precision of survey estimates for subpopulations important to the survey's objectives. In the 2022-2023 NSFG, those groups included teenagers (15-19 years of age), non-Hispanic Black persons, and women.

Target population—The population to be described by estimates from the survey. In the NSFG the target population is the household population of the United States, which refers to the civilian

noninstitutionalized population, plus active-duty military who are not living on military bases. "Noninstitutionalized" refers to the omission of prisons, hospitals, dormitories, and other large residences under central control. College students living in dormitories were included but sampled through their parents'/guardians' households.

Trimming—The process of reducing very large weights for individual cases in the data set. Trimming may be done to reduce the effects of very large individual weights on sample statistics, to reduce disclosure risks from such large weights, and to reduce potential bias in statistics resulting from these very large weights. Trimming occurs during the last stage in the process of creating sampling weights.

Weight—See "Sampling Weight."