Public-Use Data File Documentation

2022-2023
National Survey of Family Growth

USER'S GUIDE

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
Centers for Disease Control and Prevention
National Center for Health Statistics

Hyattsville, Maryland
December 2024

TABLE OF CONTENTS

NSFG Website

Data users can obtain the latest information about the National Survey of Family Growth (NSFG) by periodically checking our website: https://www.cdc.gov/nchs/nsfg. The website features downloadable data and documentation for the 2022-2023 NSFG and previous file releases, as well as important information about any modifications or updates to the data or documentation. Published reports from previous surveys are also available, as are updates about future surveys and datasets. Data files and documentation can be found at: https://www.cdc.gov/nchs/nsfg/nsfg_questionnaires.htm

NSFG Announcements Listserv

Data users are encouraged to join the NSFG Announcements Listserv, an electronic mailing list. The Listserv is composed of NSFG data users located around the world who receive news about the survey (e.g., new releases of data) and publications. To join, go to https://www.cdc.gov/nchs/products/nchs_listservs.htm, and select "National Survey of Family Growth (NSFG) Announcements" to subscribe.

Questions about NSFG?

Most commonly asked questions about the NSFG are addressed in the "Frequently Asked Questions" posted on the NSFG webpage and throughout this User's Guide. If, however, you have reviewed this User's Guide and the other materials posted on the NSFG webpage thoroughly and you still have a question, please contact the NSFG team at nsfg@cdc.gov.

Guidelines for Citation of Data Source

With the goal of mutual benefit, the National Center for Health Statistics (NCHS) requests that recipients of NSFG data files cooperate in certain actions related to their use. Any published material derived from the 2022-2023 NSFG data should acknowledge "National Center for Health Statistics, National Survey of Family Growth" as the original source. The full spelling of the source without the use of acronyms is preferred.

The suggested citation to appear at the bottom of all tables and graphs is as follows:

Data Source: National Center for Health Statistics, National Survey of Family Growth, 2022-2023.

In a bibliography, the suggested citation for this document is:

> National Center for Health Statistics. National Survey of Family Growth, 2022-2023 User's Guide. 2024. Available from: http://www.cdc.gov/nchs/nsfg/nsfg-2022-2023-puf.htm

The suggested citation for 2022-2023 NSFG survey data and other documentation is:

> National Center for Health Statistics. National Survey of Family Growth, 2022-2023. Public use data file and documentation. http://www.cdc.gov/nchs/nsfg/nsfg-2022-2023-puf.htm

The published material should also include a disclaimer that credits the author's analyses, interpretations, and conclusions to the author (user of the data file) and not to NCHS, which is responsible only for the initial data release. Users who wish to publish a technical description of the data should make a reasonable effort to ensure that the description is consistent with that published by NCHS.

NSFG questionnaires are in the public domain and no permission is required to use them. Citation as to source, however, is appreciated.

Information on how to cite NCHS publications and electronic media is available at: https://www.cdc.gov/nchs/products/citations.htm.

DATA USER'S AGREEMENT

The National Center for Health Statistics (NCHS), Centers for Disease Control and Prevention (CDC), conducts statistical and epidemiological activities under the authority granted by the Public Health Service Act (42 U.S.C. § 242k). NCHS survey data are protected by Federal confidentiality laws including Section 308(d) Public Health Service Act [42 U.S.C. 242m] and the Confidential Information Protection and Statistical Efficiency Act or CIPSEA (Title III of the Foundations for Evidence-Based Policymaking Act of 2018 (Pub. L. No. 115-435, 132 Stat. 5529 § 302)). These confidentiality laws state the data collected by NCHS may be used only for statistical reporting and analysis. Any effort to determine the identity of individuals and establishments violates the assurances of confidentiality provided by federal law.

Terms and Conditions

NCHS does all it can to assure that the identity of individuals and establishments cannot be disclosed. All direct identifiers, as well as any characteristics that might lead to identification, are omitted from the dataset. In addition, some records have had one or more responses changed through statistical perturbation. These modifications are intended to prevent definitive identification of individual respondents and their responses. They do not affect univariate point estimates and have a minimal effect on estimates of variance and tests of statistical significance. Any intentional identification or disclosure of an individual or establishment violates the assurances of confidentiality given to the providers of the information. Therefore, users will:

1. Use the data in this dataset for statistical reporting and analysis only.

2. Make no attempt to learn the identity of any person or establishment included in these data.

3. Not link this dataset with individually identifiable data from other NCHS or non-NCHS datasets.

4. Not engage in any efforts to assess disclosure methodologies applied to protect individuals and establishments or any research on methods of re-identification of individuals and establishments.

By using these data you signify your agreement to comply with the above-stated statutorily based requirements.

Sanctions for Violating NCHS Data Use Agreement

Willfully disclosing any information that could identify a person or establishment in any manner to a person or agency not entitled to receive it, shall be guilty of a class E felony and imprisoned for not more than 5 years, or fined not more than $250,000, or both.

ONE-PAGE SUMMARY FOR USERS OF THE 2022-2023 NSFG PUBLIC-USE DATA

This User's Guide provides general information for users of the public-use data, including:

- An overview of the 2022-2023 NSFG and what is new in the survey design since previous surveys
- How to access NSFG data and documentation files
- How the NSFG data files are organized
- Information on sample weights and variance estimation
- How the NSFG survey data were prepared for public use
- Descriptions of the codebooks and questionnaires

The Frequently Asked Questions (FAQ) about the NSFG, posted on the NSFG webpage, would be a good starting point for current or prospective users of the public-use data and documentation files. Here are the primary highlights from the FAQ:

- **Use sample weight (e.g., WGT2022_2023) and design variables (VECL & VEST)** to make valid estimates. Failure to use the weights and design variables correctly will lead to inaccurate statistical estimates and inferences.

- **Use recoded variables when available.** The recode variables have been evaluated carefully, and values have been imputed for many of these recodes where the source variables have missing data. Some of the most commonly used recodes are listed on page 24.

- Everything users need to conduct analyses with the 2022-2023 NSFG public-use data is provided on the **NSFG website**:
    - Downloadable **CSV and SAS data files**
    - **Codebooks** with entries for each variable on the public-use files
    - **File indexes** listing all variables on the public-use files
    - **Recode specifications** for specially constructed variables, many of which were imputed on missing values
    - **Questionnaires** in two levels of detail
    - Summary of NSFG **questionnaire changes since 2017-2019**
    - **Topic-specific notes** for analysis
    - **Variance estimation examples** using 2022-2023 data that can be adapted for your own research purposes
    - Lists of **restricted-use analytic and contextual variables** available through NCHS RDC

Please refer to the sections below on "**Survey and Sample Design for 2022-2023; New Features**" and "**Sample Weights and Variance Estimation**" for information on the sample design, data collection plan, and other procedures of the survey until further information is published on the NSFG webpage specific to the 2022-2023 survey design and operations.

BACKGROUND AND OVERVIEW OF THE 2022-2023 NSFG

The National Survey of Family Growth (NSFG) is designed and administered by the National Center for Health Statistics (NCHS), an agency within the U.S. Department of Health and Human Services' Centers for Disease Control and Prevention (DHHS/CDC). NCHS conducts the NSFG in collaboration with several other agencies of the DHHS. (See **Acknowledgments** for further detail on these cosponsoring agencies.)

The NSFG became part of the federal statistical system, within NCHS, in 1973. The primary purpose of the survey, particularly since the inclusion of a sample of men as well as women aged 15-44 (15-49 starting in September 2015), has been to produce reliable national estimates of:

- factors affecting pregnancy and live birth, including sexual activity, contraceptive use, and infertility
- medical care associated with contraception, infertility, and childbirth
- factors affecting marriage, divorce, cohabitation, and family building
- adoption and caring for non-biological children
- father involvement with their children
- use of sexual and reproductive health services
- attitudes about sex, childbearing, and marriage
- risk of HIV and STIs in general household population

The following table presents basic information on each NSFG public-use file release since 1973.

Cycle	Year	Scope	Number of Respondents	Over-Samples	OMB Approved Length	Respondent Incentive	Response Rates
1	1973	Ever-Married Women 15-44	9,797	Black Women	60 Minutes	No	90.2%
2	1976	Ever-Married Women 15-44	8,611	Black Women	60 Minutes	No	82.7%
3	1982	All Women 15-44	7,969	Black Women Teens	60 Minutes	No	79.4%
4	1988	Women 15-44	8,450	Black Women	70 Minutes	No	82.5%
5	1995	Women 15-44	10,847	Black Women Hispanic Women	100 Minutes	$20	78.7%
6	2002	Women 15-44 Men 15-44 (First time)	12,571 W = 7,643 M = 4,928	Black persons, Hispanic persons, 15-24 year olds	W= 85 min M= 60 min	$40	79% W=80% M=78%
n/a	2006-2010	Women 15-44 Men 15-44	22,682 W = 12,279 M = 10,403	Black persons, Hispanic persons, Teens	W=80 min M=60 min	$40	77% W=78% M=75%
n/a	2011-2013	Women 15-44 Men 15-44	10,416 W = 5,601 M =4,815	Blacks Hispanics Teens	W=80 min M=60 min	$40	72.8% W=73.4% M=72.1%
n/a	2013-2015	Women 15-44 Men 15-44	10,205 W=5,699 M=4,506	Blacks Hispanics Teens	W=80 min M=60 min	$40	69.3% W=71.2% M=67.1%
n/a	2015-2017	Women 15-49 Men 15-49	10,094 W=5,554 M=4,540	Black persons Hispanic persons Teens	W=80 min M=60 min	$40	65.3% W=66.7% M=63.6%
n/a	2017-2019	Women 15-49 Men 15-49	11,347 W=6,141 M=5,206	Black persons Hispanic persons Teens	W=80 min M=60 min	$40	63.4% W=65.2% M=61.4%
n/a	2022-2023	Women 15-49 Men 15-49	9,957 W=5,586 M=4,371	Black persons Teens	W=75 min M=50 min	$40	23.4% W=23.8% M=23.1%

WHERE TO FIND NSFG PUBLIC-USE DATA AND DOCUMENTATION and OTHER NSFG DATA FILES

The public-use data and documentation for the 2022-2023 NSFG are available on the NSFG website. The NSFG public-use data are contained in three downloadable data files: female respondent, female pregnancy, and male respondent. See the section "**Organization of the 2022-2023 NSFG Public-Use Data Files**" for more details.

NSFG documentation for each data file consists of:

- Codebooks, showing separate entries for each variable on the public-use files
- File indexes, listing all variables included on the public-use files
- Recode specifications, describing constructed variables available on each file
- Lists of restricted-use analytic variables available through the NCHS Research Data Center, with submission of an approved research proposal.

In addition to this User's Guide and the file-specific documentation noted above, users are provided with the following resources on the NSFG webpage:

- Frequently Asked Questions about the NSFG
- Guidance on Variance Estimation, Sample Design, and Weighting
- Topic-Specific Notes
- Summary of Questionnaire Changes Since 2017-2019
- Female and Male Questionnaires for 2022-2023 NSFG
- Other Survey Implementation Materials for 2022-2023

NEW FOR THE 2022-2023 NSFG

- Data collection for the 2022-2023 NSFG was conducted using a multimode design, including traditional face-to-face (FTF) interviewing for some respondents and web survey administration for others. This transition from FTF-only to a multimode design was motivated primarily by the challenges posed by the COVID-19 pandemic. While a significant shift from past data releases, the multimode design used in 2022-2023 maintained key features of the survey design and operation of past NSFG releases since 2006. See the section further below on "**Survey and Sample Design for 2022-2023**" for further details.

- The survey questionnaires for 2022-2023 were streamlined to reduce respondent burden, help minimize potential for disclosure risk, and simplify the organization of the data files for users. See **Summary of NSFG Questionnaire Changes Since 2017-2019** posted on the NSFG webpage for further details.

- There are a number of new variable suppressions and modifications related to disclosure risk reduction for the 2022-2023 public-use files. See the section further below on **"Protections**

to Minimize Risk of Disclosure of Individual-Level Data" as well as the lists of restricted-use analytic variables posted on the NSFG webpage for more information.

ORGANIZATION OF THE 2022-2023 NSFG PUBLIC-USE DATA FILES

The public-use data for the 2022-2023 NSFG are provided as three separate data files in CSV format and SAS7BDAT format, as described below.

DATA FILE	**Number of Records (observations)**	**Number of Variables**
Female respondent file File = 2022_2023_FemRespData.csv (and *.sas7bdat) (one record per female respondent)	5,586	1,912
Female pregnancy (interval) file File = 2022_2023_FemPregData.csv (and *.sas7bdat) (one record per pregnancy reported by female respondents)	8,247	111
Male respondent file File = 2022_2023_MaleData.csv (and *.sas7bdat) (one record per male respondent)	4,371	1,157

"Read-me" files are included with each CSV data file download that provide additional information for importing the data files into SAS, Stata, and R. Data users can refer to the file indexes and codebooks for variable and value labels. The sasfiles provided include variable labels.

Data Layout for Each Public-Use File

The following is a listing of the major sections in the three public-use files from the 2022-2023 NSFG. Not all items asked in the NSFG questionnaires could be included on the public-use files due to disclosure risk concerns. Please see section on **Protections to Minimize Risk of Disclosure of Individual-Level Data** for further details on actions taken to reduce disclosure risk with these data files. In addition, we note that the survey questionnaires for 2022-2023 were also streamlined to reduce respondent burden and help minimize potential for disclosure risk. These changes include revising questions with continuous responses to have categorical response options and reducing the number of date questions asked of respondents. Some detail previously collected was reduced or restructured to simplify file structure, notably for the male file. See **Summary of NSFG Questionnaire Changes Since 2017-2019** posted on the NSFG webpage for further details on how the male and female questionnaires were modified.

File Indexes for each of the public-use files provide more detail, including short labels

and variable types for every variable included on the public-use files, as well as an asterisk to indicate those variables where some modification was made for disclosure risk reduction. **The restricted-use variables lists** include all restricted-use analytic and contextual variables available only through the NCHS Research Data Center.

FEMALE RESPONDENT FILE –
(1 record for each of the 5,586 female respondents in 2022-2023)

- Respondent ID Number (randomized, with no linkage to geography) and selected household screener variables
- Raw and Blaise-computed variables from the female questionnaire sections A-J
    A. Demographic Characteristics; Household Roster Summary Variables (not full roster); Childhood Background
    B. Pregnancy & Birth History; Adoption & Nonbiological Children
    C. Marital & Relationship History; 1st Sexual Intercourse
    D. Sterilizing Operations and Impaired Fecundity
    E. Contraceptive History and Pregnancy Wantedness
    F. Family Planning and Medical Services
    G. Desires and Intentions for Future Births
    H. Medical Help to Have a Baby; General & Reproductive Health
    I. Health insurance; Residence and place of birth; Religion; Past and Current work (R and current spouse/partner)
    J. Computer-Assisted Self-Interview (CASI) – Health status; body mass index; opposite-sex and same-sex behavior; sexual orientation and attraction; STD/HIV risk behaviors; adverse and positive childhood events; household income and sources of income; COVID-19 experience
- Recodes and Imputation Flags based on sections A-J
  (including selected pregnancy variables shown for each woman, if she had any pregnancies; imputation flags show whether the value of a recode variable was imputed for that case)
- Weights and variables for variance estimation
- Date of interview (century month) and related variables

FEMALE PREGNANCY (INTERVAL) FILE –
(1 record for each of the 8,247 pregnancies reported by female respondents in 2022-2023)

- Respondent Case ID Number (randomized, with no linkage to geography)
- Pregnancy Order (1st, 2nd, 3rd, etc.)
- Raw and Blaise-computed pregnancy-specific variables from sections B and E
    - Section B items provide various details about the pregnancy
    - Section E items describe contraceptive use and wantedness at time of conception, as well as marital/cohabiting status at conception and outcome of pregnancy
- Recodes for section B (pregnancy descriptors such as outcome, year when began & ended, categorical version of gestational length, duration of breastfeeding, etc.)
- Recodes for section E (wantedness, marital/cohabiting status at pregnancy start & end)
- Respondent characteristics (recodes)

- Each pregnancy record includes some respondent-based variables for demographic characteristics, (e.g., race, religion, education at interview) provided for user convenience.
- Imputation flags for each recode included on pregnancy file
- Weights and variables for variance estimation
- Date of interview (century month) and related variables

MALE RESPONDENT FILE –
(1 record for each of the 4,371 male respondents in 2022-2023)

- Respondent ID Number (randomized, with no linkage to geography) and selected household screener variables
- Raw and Blaise-computed variables from questionnaire sections A-Ks
  A. Demographic Characteristics; Household Roster Summary Variables (not full roster); Childhood Background
  B. Ever Sex with a Female, Sex Communication and Education, Vasectomy and Physical Ability to Father Children, Number of Female Sexual Partners, Enumeration and Relationship With Up To 3 Recent (Or Last) Female Sexual Partner(s)
  C. Current marriage or cohabitation
  D. Recent (or last) female sexual partners
  E. First Former Wife; First Female Cohabiting Partner; First Female Sexual Partner
  F. Biological Children Ever Fathered; Nonbiological Children Living with R; Other Pregnancies Fathered
  G. Parenting activities with coresidential and non-coresidential children
  H. Desires and intentions for future biological children
  I. Health insurance; Health conditions and health services
  J. Residence and place of birth; Religion; Past and Current work (R and current spouse/partner)
  K. CASI: Health status; body mass index; opposite-sex and same-sex behavior; sexual orientation and attraction; STD/HIV risk behaviors; adverse and positive childhood events; household income and sources of income; COVID-19 experience
- Recodes and Imputation Flags based on sections A-K
- Weights and variables for variance estimation
- Date of interview (century month) and related variables

SURVEY AND SAMPLE DESIGN FOR 2022-2023

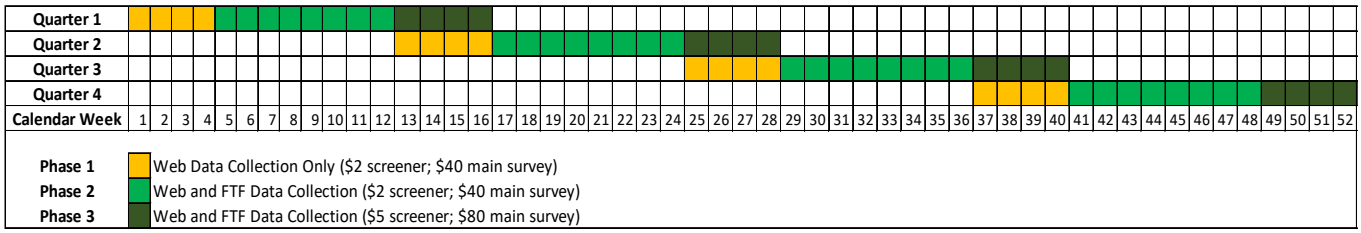
Survey Design

The 2022-2023 NSFG survey design includes elements that are consistent with prior NSFG data collection, as well as important changes. The most significant change involves going from solely FTF interviewing with an audio CASI (ACASI) portion at the end, to data collection being conducted with online, or web, surveys for some respondents and FTF interviews with a

text-only (no audio) CASI portion at the end, for others. This new design is referred to as "multimode." The use of multimode was accelerated to begin at the launch of data collection in January 2022 rather than having a gradual transition after the first 2 years of data collection informed by a mode experiment as originally planned, primarily due to the COVID-19 pandemic prohibiting FTF interviewing. Consistent with successful features of the NSFG surveys that took place from 2002 to 2019, the 2022-2023 design also used multiple phases of data collection with subsampling and enhanced protocols for a later phase characterized by higher incentives, but with the addition of an initial web-only phase. This multimode, multiphase, quarterly design is depicted in the figure below across all 52 weeks of a year. The 2022-2023 design includes three phases:

- **Phase 1**: In the first four weeks of each quarter, data were collected only with web surveys.
- **Phase 2**: During phase 2, beginning in the 5th week of each quarter and lasting 8 weeks, FTF data collection took place for the sample members who had not yet completed the web survey. Web data collection remained an option as well.
- **Phase 3**: For the final 4 weeks of each quarter (e.g., weeks 13-16 for quarter 1), data collection continued for only a subsample of screener and main survey nonrespondents, as well as all respondents who started but did not finish the main survey, using increased incentives and both modes (FTF and web) for screener and main survey completion. Consistent with previous NSFG surveys since 2002, this enhanced protocol for a subset of nonrespondents has been effective for controlling costs and reducing potential nonresponse bias. Screener and main survey data collection for each quarter concluded at the end of week 16 for that quarter. Since there are 12 weeks in a calendar quarter, and a total of 16 weeks for phases 1, 2, and 3, the last 4 weeks of a quarter overlap with the first 4 weeks of the next quarter. For example, the figure below shows that weeks 13-16 of the year include both Phase 3 of Quarter 1 (dark green blocks) and Phase 1 of Quarter 2 (yellow blocks). During 2022-2023 data collection, 74% (7,375) of main surveys were completed in web mode and 26% (2,582) in FTF mode.

Multimode, Multiphase, Quarterly Design



Sample Design

The 2022-2023 NSFG sample design, in many aspects, was consistent with the design implemented for the NSFG for 2011-2019 (see: "2017–2019 National Survey of Family Growth (NSFG): Summary of Design and data collection methods" and the analogous document for 2022-2023, for more details on similarities and differences). The 2022-2023 NSFG design was based on a national area multi-stage probability sample with a four-stage clustered sample design, with a 5th stage for nonresponse follow-up. The five stages below are generally the same

as those used in the 2017-2019 and prior sample design, thus preserving basic sampling methodology, but with differences in the details within the stages. These changes reflect efforts to improve efficiency, reduce variation, and reduce costs within the stages.

- The first stage involved the selection of Primary Sampling Units (PSUs). There were 80 PSUs included in the sample across the 2 years, and 40 in each year, with the latter comprised of 25 PSUs that were only included in one year, plus 15 PSUs that were included in both years.

- The second stage involved selection of SSUs or segments within PSUs. These are defined as Census Block Groups (CBGs) or groupings of two or more adjacent CBGs in cases where the expected number of eligible households was insufficient to support fieldwork. Twelve SSU were sampled from each PSU each year. To achieve an oversample of Black respondents, the chance of selection of SSUs with higher percentage of Black individuals (based on most recently available ACS figures) was increased relative to other CBGs.

- The third stage selected households within SSUs using the contractor's in-house enhanced address-based sampling frame (ABS) and field-enumerated segments. The households in ABS segments were stratified by the likelihood of containing one or more age-eligible individuals, and households were sampled at a higher rate from the higher likelihood strata. There was no stratification of households in field-enumerated segments.

- The fourth stage involved selecting one of the eligible persons within a household. The within-household selection rates were set so that about 20% of all main surveys were with teens aged 15-19 and 55% of all main surveys were with females.

- A fifth stage of sampling involved the double sampling of nonrespondents during Phase 3 of data collection within each quarter, described above in "Survey Design."

During 2022-2023 data collection, 98,307 households were sampled, and 28,505 households completed screener surveys with the multimode design, yielding 9,957 completed main surveys (5,586 females and 4,371 males). Each single year of data is designed to be nationally representative, however, users of the 2022-2023 NSFG should use both years of data to permit statistically reliable estimates to be made. Weights are provided for the two-year dataset.

The data included in this 2022-2023 NSFG public-use file release are nationally representative of the household population aged 15-49. Data collection for the 2022-2023 NSFG was conducted from January 2022 through December 2023, based on survey protocol and informed consent procedures approved by the NCHS Ethics Review Board (protocol #2021-07). One sample respondent per household was selected based on a short household screener to determine eligibility and with the selection algorithm for oversamples in place as described above. Interviews taking place in FTF mode were conducted by female interviewers trained specifically for the NSFG survey using tablets programmed with the survey questionnaires (computer-assisted personal interviewing or CAPI), with a self-administered component at the

end. Electronically signed parental permission and minor assent were obtained for all minor respondents. Adult respondents for the 2022-2023 NSFG, as in prior survey years, could provide their consent verbally without signature. In 2022-2023, the completed surveys for female respondents averaged 75 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)

The following paragraphs provide, first, response rates and summaries of the response rate calculation methodology, and second, basic information on data quality. More details on these topics and other methodology topics can be found in the following 4 reports based on the 2022-2023 NSFG (expected to be released by summer 2025):

- Summary of Design and Data Collection Methods
- Sample Design Documentation
- Sample Error Estimation Design
- Weighting Design Documentation

Calculation of Response Rates (Summary)

The basic formula for calculating the overall response rate is: (Screener RR) x (Main Survey RR). As described in *Survey Design* above, the design includes 3 phases, with the third phase involving subsampling (or a double sample) of the remaining non-responding sample. Thus, the response rate calculation can be understood as a double-sample response rate calculation, similar to that used for previous NSFGs in 2017-2019 and prior (see "2017-2019 National Survey of Family Growth (NSFG): Sample Design Documentation"). Both screener and main survey response rates are calculated from: the 12-week period (Phase 1 below) covered by the first two phases of quarterly sample release; and the 4-week period (Phase 2 below) after subsampling of nonrespondents, as follows:

$$(\text{Phase 1 RR}) + [1 - (\text{Phase 1 RR})]*(\text{Phase 2 RR})^{1}$$

Further details for 2022-2023 NSFG will be available by Summer 2025 in the 4 methodology reports listed above, including methodology for estimating the eligibility rate of households for the screener response rate, and methodology for applying base weights to response rates to obtain weighted response rates.

[1] https://aapor.org/wp-content/uploads/2023/05/Standards-Definitions-10th-edition.pdf. The American Association for Public Opinion Research. 2023 Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 10th edition. AAPOR.p. 90.

Response Rates

The below table shows 2022-2023 response rates (percentages) for the screener survey, the main survey, and overall, meaning final response rates (RRs) combining screener and main survey RRs.

Weighted* response rates: 2022-2023			
	Screener	Main	**Combined**
Total	42.2	55.5	23.4
Total Females	42.2	56.3	23.8
age 15-19	42.2	51.2	21.6
age 20-49	42.2	57.1	24.1
Total Males	42.2	54.8	23.1
age 15-19	42.2	51.0	21.5
age 20-49	42.2	55.4	23.4

*Base weights applied. All response rates reflect adjustment for the double sample, in other words, phase 3.

Data Quality, Nonresponse Bias Analysis Summary, and Information Related to Response Rates

Response rates in household surveys have been declining, and this trend has been observed on NSFG, even prior to 2022-2023 NSFG. The NSFG response rates for 2022-2023 are markedly lower that the response rates for the last 2-year data collection period of 2017-2019. Some of the sharper decline in participation can be attributed to the effects on society of the COVID-19 pandemic, from respondents' willingness to participate to the ability to recruit and retain interviewers. Complex and long FTF surveys such as the National Health and Nutrition Examination Survey (NHANES) that did not undergo major design changes from before the pandemic, also experienced drops in participation. For example, the NHANES unweighted interview response rate fell from 51.0%[2] for the 2017-2020 data collection to 34.5%[3] for 2021-2023. The examination response rates fell from 46.9% to 25.6%. Other similarly long and sensitive household surveys to the NSFG, such as the National Survey of Drug Use and Health (NSDUH) that underwent a redesign from only FTF to a web and FTF design, like NSFG, have also observed substantial drops in survey participation. The weighted overall survey response rate for NSDUH dropped from 45.8%[4] in 2019 (pre-pandemic) to 12.3%[5] in 2023.

Long and complex surveys such as NSFG pose challenges for web administration, with approaches still being developed and evaluated to increase participation. During the 2022-2023 NSFG data collection, a responsive design was employed testing and implementing a number of

[2]https://wwwn.cdc.gov/nchs/data/ResponseRates/NHANES-2017-2020-Response%20Rates-2017-March2020.pdf
[3]https://wwwn.cdc.gov/nchs/data/ResponseRates/NHANES-August-2021-August-2023-Response-Rates.pdf
[4]https://www.samhsa.gov/data/sites/default/files/reports/rpt29395/2019NSDUHMethodsSummDefs/2019NSDUHMethodsSummDefs082120.pdf
[5]https://www.samhsa.gov/data/sites/default/files/reports/rpt47098/Methodological%20Summary%20and%20Definitions/2023-nsduh-method-summary-defs.pdf

improvements to increase participation, including the design of materials, additional ways to log into the survey, and increased incentive amounts. Some changes were tested and ultimately not implemented, such as the use of a mailed paper screener. Lastly, some changes were made without testing, such as improvements to the instrument to decrease survey breakoffs, addition of ways to send survey reminders, and improving strategies to follow up with parents to obtain consent for minors.

The responsive design aimed at reduction of the risk of nonresponse bias worked as intended. Evaluations of the increased incentive amount starting in Phase 1 showed significantly higher response rates, improved demographic representation, and impact on some survey estimates. Evaluation of Phase 3 showed that those who did not participate until being offered a higher incentive in Phase 3, were significantly different from the Phase 1 and 2 respondents on multiple key survey estimates. Experiments in survey nonresponse have shown that incentives can help gain participation from people who are less interested in the survey topic and can reduce bias in survey estimates[6][7][8]. Recruiting respondents in Phase 3 who are different provides further evidence in support of the responsive design and its role in reducing the risk of nonresponse bias.

Data users are encouraged to exercise caution when making comparisons to prior years of data collection and interpreting any differences. The multimode design is very different from the prior FTF-only design, and the COVID-19 pandemic prevented experimental evaluation of the impact of the multimode design on survey estimates. Like other surveys that had to change to a new design due to the disruption of FTF interviewing, changes in estimates from before and after the pandemic are confounded by real changes in the population, measurement differences, and nonresponse, among other possible sources of error (coverage, sampling, and processing). As a result, some estimates may show larger differences from the 2017-2019 NSFG and earlier data compared to those observed between prior data releases .Users are advised to note the design-related changes, particularly with regard to mode, if making statements comparing earlier estimates to those from 2022-2023. Further evaluations, including nonresponse bias analysis, will be included along with the detailed methodology documents listed in the *Sample Design* section.

SAMPLE WEIGHTS AND VARIANCE ESTIMATION

Since the NSFG is a multi-stage probability-based, nationally representative sample of the household population aged 15-49, and not a simple random sample of the population, data users should understand how to account for the complex sample design when doing their analyses in order to obtain statistically valid results. This section provides a summary of the procedures used for sample weighting and variance estimation. More detailed information on the

[6]Groves, R. M., Singer, E., & Corning, A. (2000). Leverage-saliency theory of survey participation - Description and an illustration. Public Opinion Quarterly, 64(3), 299-308. Retrieved from <Go to ISI>://000166024000004

[7] Groves, R. M., Presser, S., & Dipko, S. (2004). The role of topic interest in survey participation decisions. Public Opinion Quarterly, 68(1), 2-31. Retrieved from <Go to ISI>://000221299300001

[8] Groves, R. M., Couper, M. P., Presser, S., Singer, E., Tourangeau, R., Acosta, G. P., & Nelson, L. (2006). Experiments in Producing Nonresponse Bias. Public Opinion Quarterly, 70(5), 720-736.

2022-2023 sample weights will be provided in the methodology documentation to be available in Summer 2025, see "Sample Design" section.

Sample Weights

Each respondent in the 2022-2023 NSFG sample represents a different number of people in the U.S. household population, and this number is indicated in the respondent's sample weight. There are several factors that lead to variation in the size of the weights. For example, Black persons and teens were selected at higher rates than others in the 15-49 age group. Women also had a slightly higher probability of selection than men. Sample weights adjust for these unequal probabilities of selection for different population subgroups. The sample weights were further adjusted to account for differential response rates and coverage rates, so that accurate national estimates can be made from the sample. The weights were calibrated to the ACS 2022 1-Year PUMS File. The calibration step included sex, age, race/ethnicity, highest education level attained, and marital status. Data users should use the weights in all analyses to obtain accurate, statistically valid estimates. Using the weights will also permit replication of the nationally representative estimates that appear in published NCHS reports.

Each of the 2022-2023 data files have a weight variable called "WGT2022_2023" with values for each of the 5,586 female and 4,371 male respondents who completed NSFG survey in 2022-2023. When correctly applied for the full set of cases, this "WGT2022_2023" variable yields estimates representative of the 74.9 million women and 75.7 million men in the household population aged 15-49 of the United States in 2022.

To yield the population number in thousands, as often appears in NCHS reports, you would divide the sample weight by 1,000. For example, you could create a new weight variable as shown below:

WGT1000=WGT2022_2023/1000

In addition to using the sample weight variable WGT2022_2023, researchers must use the design variables for the sampling stratum (VEST) and cluster (VECL) to obtain correct standard errors for their estimates.

Variance Estimation

Sampling variance is a measure of the precision of a statistic (such as a percentage, proportion, or a mean) due to having taken a sample of instead of interviewing or measuring all members of the full population. In the 2022-2023 NSFG, the sampling variance measures variation from the full population-based parameters due to interviewing the NSFG sample of 9,957 respondents instead of all 150.6 million women and men aged 15-49 in the US household population.

Many statistical software packages by default compute "population" variances, which may underestimate the sampling variances because they assume that the sample was drawn using simple random sampling. Statistical software that can analyze data drawn from complex survey

samples is required to accurately estimate sampling errors in a complex sample such as the NSFG. For example, SAS procedures such as "FREQ" produce population variances assuming simple random sampling, but SAS has procedures for complex survey estimates in its 'SURVEY' procedures such as "SURVEYFREQ." Similarly, SUDAAN, Stata, R, and SPSS have procedures designed to analyze data derived from a complex sample survey.

When estimating variances for population subgroups (such as those who have ever had sexual intercourse or those 20-49 years of age), it is important to read in the entire data set first. An indicator variable for your subpopulation (e.g., in SAS for 20-49 year olds, "if AGER >= 20 then agepop=1") should be created to identify only those observations that will be used in the analysis. Then, define your subgroup of interest within your survey procedure, such as using the SUBPOPN command in SUDAAN, the SUBPOP command in Stata, or by using the DOMAIN statement in SAS or including the variable in your SURVEYFREQ table command. If the data are subset without first reading in the entire data set, then empty clusters may be lost, making the survey design structure incomplete. This may result in the statistical software terminating or leading to other errors. Three **variance estimation examples** using SAS and Stata for the 2022-2023 file are posted on the NSFG webpage. Example 3 shows how to create a subpopulation variable in SAS and Stata. Further examples using R and SUDAAN will be made available later in 2025.

OVERVIEW OF DATA QUALITY IN THE NSFG

High quality data was obtained for the 2022-2023 NSFG through:

- Questionnaire design work, including careful specification, testing, and incorporation of lessons learned from the past NSFG data collection periods;
- Simplification and adaptation of the survey instruments and aids to work well in web mode where no field interviewer can facilitate the survey completion;
- Consistency checks built into the survey that allowed potential data problems to be resolved in the field rather than after data collection;
- Visual aids for face-to-face mode include showcards presenting response categories and a paper calendar for accuracy in remembering dates of events and recording them in correct chronological order; in web mode, an electronic version of the calendar is presented to respondents with events automatically populated after they answer questions in the survey.
- Evaluation of periodic data files to find and correct instrument problems before significant numbers of cases were affected; and
- Extensive interviewer training to ensure adherence to consistent and ethical fieldwork procedures.

Data files as large and complex as these cannot be guaranteed to be free of errors. If you believe you have found an error or need further assistance that cannot be found in materials provided on the webpage, please email the NSFG staff at NCHS at **nsfg@cdc.gov.**

DATA PREPARATION FOR PUBLIC USE

This section describes steps taken to prepare the NSFG survey data for public use. Some of these actions were taken simply to make the data more accurate and useful. Other actions were taken to protect the confidentiality of individual respondents, in keeping with the legal and ethical obligations of NCHS when conducting the NSFG or any of its other surveys.

Logical Inconsistencies and Out-of-Range Values

During data collection, logical consistency across data items was maintained through "edit checks" built into the male and female survey instruments. In FTF mode, these edit checks relied on the interviewers to facilitate resolutions by the respondent during the interview. As part of the adaptation of the NSFG survey instruments to work effectively for web respondents, these edit checks were made more manageable for respondents to navigate on their own. Regardless of mode, these edit checks alerted the interviewer (or web respondent) to inconsistent or out-of-range entries and required some attempt to correct the entry. Out-of-range values are minimal in the 2022-2023 NSFG public-use files (as in past data releases) because valid ranges were specified and programmed into the survey instrument to the extent possible, and values outside that range were rejected or signaled as something to correct.

Some edit checks in the instrument are **"hard edits"** in that they disallow combinations of values that are impossible (for example, respondents cannot report a date for any event in their lives that is later than the interview date or before their date of birth). Other edit checks are **"soft edits"** in that they alert the interviewer or web respondent to situations that are rare but not impossible (for example, a respondent reports that she had her first menstrual period at a particularly young age).

In soft edit checks, the respondent is given the opportunity to revise their responses in case they were given in error. If the respondent says that the information is accurate, they or the interviewer can override or move past the inconsistency warning box. For FTF interviews, the interviewers are trained to enter a brief comment to explain the situation, and these comments are reviewed to assess the need for any data edits, or possibly an enhancement of the survey instrument. In all such cases, the seemingly inconsistent data may remain on the data file. It is not possible to foresee and specify all the edit checks that might be needed in these very complex survey instruments, and as a result, some inconsistencies in the data could not be eliminated.

In addition to edit checks, other aspects of the NSFG survey instruments designed to maximize consistency *during* data collection were, for females: 1) "summary screens" before or after key sections, reminding the respondent of events and dates reported earlier, and 2) prompts to record events on, or refer to, the life history calendar mentioned above.

As in prior NSFG file releases, the process of checking for consistency within the 2022-2023 data after data collection was focused primarily on the recoded variables and variables used to construct them. These were considered to be the most critical and most frequently used variables in the files. Considerable efforts were made to detect and resolve or document inconsistencies and unacceptable codes throughout the files. However, as noted earlier, given the

size and complexity of these data files, they may not be free of inconsistent or missing responses.

Coding for "Don't Know," "Refused," and "Not Ascertained" Values

Missing data refers to responses of "don't know" or "refused" that were entered either by the respondent or the interviewer to indicate that the respondent could not or would not provide an answer to a question. "Not ascertained" refers to relatively rare instances in which a question was erroneously skipped during the survey or the reported value needed to be suppressed for other reasons. The code for "not ascertained" was generally assigned in these cases after fieldwork was completed. Only completed cases are in the files; a case was defined as being complete if the respondent answered the last applicable question before CASI (in Section I for females and in Section J for males).

Depending on the column length of the original data items:

- "Don't know" values are coded 9, 99, 999, 9999, or 99999
- "Refused" values are coded 8, 98, 998, 9998, or 99998
- "Not ascertained" values are coded 7, 97, 997, 9997, or 99997

(The codebooks only show these codes for the variable if any cases had those particular values.)

Missing data as described above is distinct from a variable that was inapplicable -- the respondent was legitimately skipped past the question (for example, respondents who had never been pregnant were not asked questions about how their pregnancies ended). For more information on determining who was asked each question, refer to the description of universe statements in the User's Guide section entitled **"Description of Codebooks"** further below or the codebook entry for particular variables. A question that was legitimately skipped or a variable legitimately not defined for a respondent will be coded as blank, and in the codebook, is indicated by a "dot" and labeled "inapplicable*"* or "sysmis."

Recoded variables may have legitimate inapplicable values, but in most instances they do not have missing data in the form of "don't know," "refused," or "not ascertained" values because these responses were imputed to a valid value. Cases that had recode values imputed because of missing information on the source variables are identified with an imputation "flag"-- a separate variable that indicates whether or not the corresponding recode was imputed (see User's Guide section on **"Recodes and Imputation"** further below, as well as Recode Specifications posted on the webpage for each data file).

Century-Month Coding for Dates

During the NSFG survey, dates of key life events were collected as month and year. For every date asked in the survey, the month and year information was converted to "century months" by subtracting 1900 from the year, then multiplying the remainder by 12, and adding the number of the month, where January = 1, February = 2, and so on.

For instance:

The century month code for October 1987 is (87 x 12) + 10 = 1054.
The century month code for January 2000 is (100 x 12) + 1 = 1201.
The century month code for July 2022 is (122 x 12) + 7 = 1471.

The century month form is convenient for computing intervals between dates (as needed for event history analyses), and subtraction yields intervals in months. With the exception of one recoded date variable (DATEUSE1 on the female respondent file) that has a leading 9 to indicate when the value was estimated, all century month date variables in the file are 4 columns long. When a month range was reported for a month question, the months shown below were consistently assigned to enable the construction of a century month value and facilitate subsequent routing through the survey:

January-March	= 1 (January)
April-June	= 4 (April)
July-September	= 7 (July)
October-December	= 10 (October)

If a respondent said "Don't Know" or "Refused" (DK/RF) when asked to report a month, the value "6" (June) was assigned for the month. If a respondent did not report a year, the century-month variable was set to 9999 for "Don't Know" or 9998 for "Refused."

The century month codes from 841 (January 1970) through 1488 (December 2023) are shown in the table below with the years from 1970 through 2023 on the vertical axis and the months on the horizontal axis. The code for a given month and year can be found by reading across the line for the appropriate year to the column headed by the appropriate month. All surveys for the 2022-2023 NSFG were conducted between January 2022 (century month 1465) and December 2023 (century month 1488).

As first done for the 2015-2017 NSFG public-use files in response to disclosure risk concerns associated with a number of dates collected in the NSFG, the public-use files for 2022-2023 no longer include all century-month dates collected in the main survey or their raw month variables. The year variables for suppressed century month dates are available for public use, but the month and century month variables are restricted to use in the Research Data Center. The section of this User's Guide **"Protections to Minimize Risk of Disclosure for Individual-Level Data"** has more information.

Century Month Codes

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1970	841	842	843	844	845	846	847	848	849	850	851	852
1971	853	854	855	856	857	858	859	860	861	862	863	864
1972	865	866	867	868	869	870	871	872	873	874	875	876
1973	877	878	879	880	881	882	883	884	885	886	887	888
1974	889	890	891	892	893	894	895	896	897	898	899	900
1975	901	902	903	904	905	906	907	908	909	910	911	912
1976	913	914	915	916	917	918	919	920	921	922	923	924
1977	925	926	927	928	929	930	931	932	933	934	935	936
1978	937	938	939	940	941	942	943	944	945	946	947	948
1979	949	950	951	952	953	954	955	956	957	958	959	960
1980	961	962	963	964	965	966	967	968	969	970	971	972

1981	973	974	975	976	977	978	979	980	981	982	983	984
1982	985	986	987	988	989	990	991	992	993	994	995	996
1983	997	998	999	1000	1001	1002	1003	1004	1005	1006	1007	1008
1984	1009	1010	1011	1012	1013	1014	1015	1016	1017	1018	1019	1020
1985	1021	1022	1023	1024	1025	1026	1027	1028	1029	1030	1031	1032
1986	1033	1034	1035	1036	1037	1038	1039	1040	1041	1042	1043	1044
1987	1045	1046	1047	1048	1049	1050	1051	1052	1053	1054	1055	1056
1988	1057	1058	1059	1060	1061	1062	1063	1064	1065	1066	1067	1068
1989	1069	1070	1071	1072	1073	1074	1075	1076	1077	1078	1079	1080
1990	1081	1082	1083	1084	1085	1086	1087	1088	1089	1090	1091	1092
1991	1093	1094	1095	1096	1097	1098	1099	1100	1101	1102	1103	1104
1992	1105	1106	1107	1108	1109	1110	1111	1112	1113	1114	1115	1116
1993	1117	1118	1119	1120	1121	1122	1123	1124	1125	1126	1127	1128
1994	1129	1130	1131	1132	1133	1134	1135	1136	1137	1138	1139	1140
1995	1141	1142	1143	1144	1145	1146	1147	1148	1149	1150	1151	1152
1996	1153	1154	1155	1156	1157	1158	1159	1160	1161	1162	1163	1164
1997	1165	1166	1167	1168	1169	1170	1171	1172	1173	1174	1175	1176
1998	1177	1178	1179	1180	1181	1182	1183	1184	1185	1186	1187	1188
1999	1189	1190	1191	1192	1193	1194	1195	1196	1197	1198	1199	1200
2000	1201	1202	1203	1204	1205	1206	1207	1208	1209	1210	1211	1212
2001	1213	1214	1215	1216	1217	1218	1219	1220	1221	1222	1223	1224
2002	1225	1226	1227	1228	1229	1230	1231	1232	1233	1234	1235	1236
2003	1237	1238	1239	1240	1241	1242	1243	1244	1245	1246	1247	1248
2004	1249	1250	1251	1252	1253	1254	1255	1256	1257	1258	1259	1260
2005	1261	1262	1263	1264	1265	1266	1267	1268	1269	1270	1271	1272
2006	1273	1274	1275	1276	1277	1278	1279	1280	1281	1282	1283	1284
2007	1285	1286	1287	1288	1289	1290	1291	1292	1293	1294	1295	1296
2008	1297	1298	1299	1300	1301	1302	1303	1304	1305	1306	1307	1308
2009	1309	1310	1311	1312	1313	1314	1315	1316	1317	1318	1319	1320
2010	1321	1322	1323	1324	1325	1326	1327	1328	1329	1330	1331	1332
2011	1333	1334	1335	1336	1337	1338	1339	1340	1341	1342	1343	1344
2012	1345	1346	1347	1348	1349	1350	1351	1352	1353	1354	1355	1356
2013	1357	1358	1359	1360	1361	1362	1363	1364	1365	1366	1367	1368
2014	1369	1370	1371	1372	1373	1374	1375	1376	1377	1378	1379	1380
2015	1381	1382	1383	1384	1385	1386	1387	1388	1389	1390	1391	1392
2016	1393	1394	1395	1396	1397	1398	1399	1400	1401	1402	1403	1404
2017	1405	1406	1407	1408	1409	1410	1411	1412	1413	1414	1415	1416
2018	1417	1418	1419	1420	1421	1422	1423	1424	1425	1426	1427	1428
2019	1429	1430	1431	1432	1433	1434	1435	1436	1437	1438	1439	1440
2020	1441	1442	1443	1444	1445	1446	1447	1448	1449	1450	1451	1452
2021	1453	1454	1455	1456	1457	1458	1459	1460	1461	1462	1463	1464
2022	1465	1466	1467	1468	1469	1470	1471	1472	1473	1474	1475	1476
2023	1477	1478	1479	1480	1481	1482	1483	1484	1485	1486	1487	1488

Recodes and Imputation

(also see File Indexes and Recode Specifications, as posted on NSFG webpage)

In order to facilitate consistent, comparable estimates of key NSFG measures for all data users, NCHS produces a number of "recoded variables," or "recodes" for each public-use file. Published NCHS reports use these recodes whenever available because they permit internally consistent and replicable estimates. NCHS also uses the recodes to prioritize the cleaning of the data file: there are too many variables in the data file to edit or reconcile them all, so NCHS focuses its cleaning and editing primarily on the recodes and on the variables that are used to construct the recodes.

Some recodes are simple, while others are complex. Some recodes may simply be transferred from single questionnaire items and imputed if missing (for example, HISPANIC, whether the respondent is of Hispanic origin). Other recodes are based on multiple questionnaire items and may involve more intricate logic to define (for example, CONSTAT1, current contraceptive status for female respondents).

Before using the original variables or constructing their own summary variables, analysts are encouraged to **check if a relevant recode exists**. Many of the raw or computed variables that have a recode corresponding to them will have a **note in the codebook** stating the name of the appropriate recode.

For convenience, below is a list of some of the more commonly used recodes corresponding to background characteristics and other key NSFG variables. Unless otherwise indicated, the recodes are available for males and females.

AGER	R's age at interview
FMARITAL	Formal (legal) marital status relative to opposite-sex spouses
RMARITAL	Informal marital status relative to opposite-sex spouses
HIEDUC	Highest completed year of school or highest degree received
HISPANIC	Hispanic origin
HISPRACE2	Race and Hispanic origin – based on 1997 OMB guidelines
INTCTFAM	Intact status of childhood family
PARAGE14	Parental living situation at age 14
EDUCMOM	Mother's (or mother figure's) education
AGEMOMB1	Age of mother (or mother figure) at first birth
METRO	Place of residence (metropolitan-nonmetropolitan)
RELIGION	Current religious affiliation
LABORFOR	Labor force status
POVERTY	Family income as percentage of federal poverty threshold
TOTINCR	Total income of R's family
HADSEX	Whether R has ever had sexual intercourse with opposite sex
VRY1STAG	R's age at first sexual intercourse
VRY1STSX	Date (century month) of first intercourse
CONSTAT1	Current contraceptive status (females only)

Besides the above list, other sources to check on the NSFG webpage are the **File Indexes** or the **Recode Specifications** to see if a relevant recode for your analyses exists.

The frequency of missing values for the recoded variables in 2022-2023 is quite low, as it was in past files. Cases that had missing data on a recode (i.e., their values could not be constructed from the source variables referenced in the recode specifications) were generally imputed, unless otherwise specified in the recode specifications.

Most missing recode values were assigned using **model-based** imputation software in which multiple regression is used to predict a value for the case using other variables in the data set as predictors. Model-based imputation follows the same logical constraints built into the recode specifications. To the extent possible, imputed values were checked to ensure that the imputed values were within acceptable ranges and were consistent with other recodes and other data reported by the respondent.

A smaller number of cases for some recodes were imputed using **logical** imputation, which involves NCHS staff examining variables related to the variable in question and assigning a value that is consistent with those other variables.

Imputation flag variables were created for every recode, allowing users to determine whether the value for each case is based on reported data, or imputed data. They also indicate which kind of imputation was used. Each imputation flag has the following potential values:

0=Questionnaire data (not imputed)
1=Multiple regression imputation
2=Logical imputation

A value of 0 on the imputation flag means that imputation was not necessary; the reported questionnaire data were sufficient to determine an appropriate value on the recode. All values other than 0 indicate that the case was imputed for this recode. The imputation process used for the 2022-2023 NSFG was similar to that used in prior releases; the **"Summary of Design and Data Collection Methods"** report on the NSFG webpage (available in Summer 2025) describes the imputation process in more detail.

As noted above, all recodes were checked thoroughly against related data items and edited if necessary, for consistency. Except when it was obviously incorrect and involved critical or commonly used variable(s), actual reported information was never replaced by an imputed value. **NCHS recommends that analysts use all cases in the file, including those with recode values imputed**. Using sample weights and including imputed cases will enable the analyst to replicate results that appear in NCHS reports. The impact of imputation on analyses can be examined by using the imputation flags to compare results with and without the imputed cases.

In addition, a subset of recode variables were defined solely on the basis of final, imputed values of other recodes, and these recodes do not have accompanying imputation flags since no imputation was needed. The specifications for these recodes are shown in blue font with a double asterisk (**) after the recode name within the **Recode Specifications** posted on the NSFG webpage.

Finding recodes in the data file and codebook: As shown in the **File Indexes** on the webpage, the recodes and their imputation flags are clustered together near the end of each of the three NSFG data files. Recodes that have accompanying imputation flags can also be distinguished from other variables by the "Variable type" displayed in their codebook entries (after the variable name and the variable label) or in the file indexes in the last column. Recodes

that do **not** have accompanying imputation flags will have variable type "computed in post-processing" to reflect their construction based on final, imputed values of other recodes.

Protections to Minimize Risk of Disclosure for Individual-Level Data

When NCHS collected data from respondents for the NSFG, those respondents were promised in the informed consent process that the information they provided would be kept confidential. NCHS is legally and ethically bound to keep that promise, both during data collection and in the production of data files for public use, while still attempting to preserve the analytic value for those who support the collection and use of these data.

As with all NCHS data files provided for public use, the proposed NSFG public-use files for 2022-2023 were submitted to the NCHS Disclosure Review Board (DRB) for review and approval. In brief, the disclosure risk protections taken for the 2022-2023 NSFG public-use files include the following, and unless otherwise indicated, replicate actions taken in earlier public-use file releases.

- All *directly* identifying information, including all names and addresses, has been eliminated from the public-use files. This information is not available within the NCHS Research Data Center (RDC).

- The only geographic variable included on the public-use files is a 3-category METRO recode (principal city of Metropolitan Statistical Area (MSA), other MSA, not MSA). All other contextual information about place of residence is only available in the RDC.

- Century month date values for key life events have been suppressed from the public-use files to prevent potential linkage or use with external data sources to identify survey respondents. These key life events include marriages, divorces, pregnancies, cohabitations, educational degrees, and selected health services.

- Other variables on the files that could potentially be used to *indirectly* identify individuals have been suppressed or modified in some way, with particular attention to keeping categories that are substantively useful and collapsing categories that were so small that they were of limited analytical use. In some cases, new variables have been created for public use based on suppressed or uncollapsed variables. For example:

    - The variable for Hispanic subgroup (HISPGRP) has been collapsed for public use, and the original variable with full detail is available only through the NCHS RDC.
    - The full household roster is only available through the NCHS RDC, and summary variables based on the household roster have been created for the public-use files.
    - The original recode PRGLNGTH indicating gestation length in weeks has again been suppressed from the public-use pregnancy file, and two categorical variables GEST_LB and GEST_OTH have been defined for public use to indicate pregnancy length for live births and other pregnancies.

- **In keeping with the changes made for the last public-use file release (for 2017-2019 NSFG), the 2022-2023 NSFG public-use files had a number of additional changes due to disclosure risk concerns.** Below is a list of the more notable changes made for 2022-2023, and the full list of all analytic variables that have been made restricted use can be found on the NSFG webpage. These restricted-use analytic variables will only be available through the RDC upon approval of a proposed research plan. *(Also see the public-use file indexes where all variables with disclosure risk reduction (DRR) actions have been asterisked, and those with new DRR actions in 2022-2023 have been highlighted in yellow)*:

    - The level of pregnancy-specific detail included on the female pregnancy file for public use has been further reduced – in particular, sex of multiple births (e.g., twins) have been suppressed and several variables such timing of first prenatal care visit have been made categorical.
    - A number of raw variables used to construct key recodes or computed variables have been suppressed for public use (for example, those related to first sexual intercourse, numbers of sexual partners, and pregnancy-specific information)
    - Several age variables that had previously been included in single years have been categorized or bottom-coded for public use (for example, ages at selected preventive health services have been top- and bottom-coded, ages at first marriage have been bottom-coded, spouse's age at marriage has been made categorical).

Whenever variables have been suppressed, modified, or newly created for reasons of disclosure risk reduction, a note has been included in the codebook. These notes are worded as follows, depending on the nature of the action taken:

For variables that have had values collapsed or categorized in some way, this note is included:

> "This variable has been modified for public use, and the original variable is accessible by application to the NCHS Research Data Center."

For century month date variables, one of the following notes is included, based on whether the CM date variable is a recode or another computed variable:

> "The month and century-month variables for this event have been suppressed for public use, but are accessible by application to the NCHS Research Data Center."

> "The year of this event is available for public use, and the original century-month variable is accessible by application to the NCHS Research Data Center."

For variables that have been created based on variables suppressed for public use, the following note is included:

> "This variable has been created for public use, and the original source variable is accessible by application to the NCHS Research Data Center."

As a final step to prevent identification of individual respondents, as done in past files, the values of some variables have been altered for a random subset of respondents through **statistical perturbation**. That is, some values in the data set are no longer the actual values reported by the respondents, resulting in greater uncertainty for anyone attempting to identify a particular individual they may know participated in the survey. However, these alterations, or statistical perturbations, were carefully designed to give analysts comparable statistical information as those obtained from the unaltered responses. In other words, it is unlikely that either national estimates or causal models are affected by any of the alterations, except for a slight increase in the variance of a few statistics.

Most of the variables suppressed from the public-use NSFG files, or variables that could not be included in their original form, are available to the research community through the NCHS RDC. The full lists of these restricted-use, analytic variables are posted on the NSFG webpage. As with all data files available through the NCHS RDC, these restricted-use data are made available to researchers under special arrangements that assure confidentiality and protection of the data. Researchers who wish to learn more about or apply for access to any of these NSFG files available through the RDC should first look at information provided on the RDC website (www.cdc.gov/rdc), and then contact either the NSFG staff at nsfg@cdc.gov or the RDC at rdca@cdc.gov.

DESCRIPTION OF CODEBOOKS

Overview

Codebooks for the NSFG provide essential information for each variable included in the public-use files. The elements of the codebook, such as variable type and universe statements, are described further below.

Below is an example page from the 2022-2023 NSFG female file codebook displaying the detailed codebook information for the female file raw variable AD-7b MARSTAT. The specific elements of the codebook are described further below.

MARSTAT	
LABEL	AD-7b R's marital or cohabiting status
VARIABLE TYPE	raw
UNIVERSE	Applicable for all respondents
NOTES	See recodes FMARITAL and RMARITAL and computed variables rmarit, fmarit, and ssmarcoh
LENGTH	8
VARNUM	9

MARSTAT: Distribution	Value	n
Married	1	2100
Living together with a partner as an unmarried couple	2	669
Neither	3	2808
Refused	8	8
Don't Know	9	1
		5586

Elements of the Codebook Entry for Each Variable

Each variable in the public-use files is represented in the codebook documentation with a page or entry containing all these elements, which are described in turn below:

- Variable name
- Variable type
- Variable label
- Universe statement
- Response categories and unweighted frequencies
- Notes, where applicable

Variable Name: For raw and computed variables, the variable name corresponds in most cases exactly to the question or computed variable name that appears in the CAPI Reference Questionnaire (CRQ). Recode and intermediate variable names correspond to those found in the recode specifications. Throughout the codebook documentation and in the recode specifications, raw and recode variables are in uppercase and computed variables are in lowercase. In some cases where questions or variables are applicable for "loops" or "arrays" (such as pregnancies, marriages, months of the year, mentions for "enter all that apply" questions, etc.), then the variable names seen in the CRQ will have numeric suffixes attached. For example, BC-8 PAYBIRTH in section B of the female CRQ is the question asking how the delivery costs were paid for this child's birth, and respondents could select all options that applied. In the 2022-2023 pregnancy file, women reported no more than 3 forms of payment for their deliveries although space was allowed for 5 mentions, so the file includes 3 variables for those 3 mentions in PAYBIRTH1-3, and this information is noted in the variable labels.

Variable Type: There are six basic variable types included in the NSFG files -- "raw," "computed," "recode," "imputation flag," "computed in post-processing," and "intermediate":

1. A **raw** variable refers to a question that was asked during the survey (the majority of variables in the data files are raw). *(In the example of PAYBIRTH1-3 above, each these 3 variables is labeled as a raw variable.)*
2. A **computed** variable is a variable computed as part of the Blaise-programmed survey instrument, based on one or more raw variables. Blaise-computed variables may play a role in subsequent routing, and their missing values are not imputed.
3. A **recode** variable is a constructed variable created, after the data are collected, from one or more raw or computed variables, and in most cases has missing values imputed.
4. An **imputation flag** is a variable indicating the mode of imputation for recodes that underwent imputation.
5. A variable **computed in post-processing** is one that was constructed after data were collected, from one or more raw variables. Like Blaise-computed variables, variables computed in post-processing do not have missing values imputed.
6. An **intermediate** variable is one that was constructed in the course of creating certain recode variables and included on the public-use files for user convenience.

Variable Label: Below each variable name in the codebook is a short variable label, identical to the label seen in the public-use file index. Variable labels differ in appearance based

on the variable type. For raw variables, each label begins with the question number as seen in the questionnaires. For example, the variables PAYBIRTH1-3 mentioned above will all show BC-8 as their question number. For computed variables (computed as part of the Blaise-programmed survey instrument), the variable label includes the "Flow Check" number from the CRQ where the computed variable was defined. For example, the variable label for male **psurgstr** shows "(Computed in FC C-12)," indicating that this variable was defined in male section C, Flow Check C-12. This particular Flow Check in the male CRQ (posted on the NSFG webpage) can be consulted to see the full specifications for how this variable was defined.

When variables are part of an array or loop, the variable labels in the codebooks (and file indexes) indicate what loop or iteration is being referenced. For example, the variable label for female CF-1 TALKPAR1-7 makes clear that each of these variables applies to the 1st through 7th mentions of sex ed topics that R has discussed with their parents.

When variables have been made categorical or top- or bottom-coded for public use, the variable labels will indicate this. This information in the label complements the codebook notes that alert the user that a variable has been modified or created for public use. For example:

- The female section A computed variable **roscnt** has this variable label, "Number of HH members based on HH roster (computed in FC A-15) (top-coded)" indicating it was top-coded for public use.
- The pregnancy file variable **AGEFATHER** has this variable label, "BC-9 Father's age when this pregnancy ended or baby/babies born (categorical)" indicating this variable was made categorical for public use.

Universe Statements ("Applicable Specifications"): In the codebook documentation, the "applicable specifications" or "universe statement" for a variable indicates which respondents were asked the question or had the variable defined for them. If a question was not applicable to a particular respondent, the questionnaire program skipped to the next applicable question. If a question was not skipped by any respondent or the variable was assigned a non-blank value for every case, the universe statement says, "Applicable for all respondents" or for the pregnancy file, "Applicable for all pregnancies." *(For example, see screen capture for AD-7b MARSTAT above.)*

Cases with inapplicable values on any variable are coded as blank or "system missing." While inapplicable values are the primary reason why variables may have blank values, they may also remain blank if a) respondents broke off before completing the survey, or b) reported values that had to be suppressed to minimize disclosure risk. Some computer programs such as SAS and Stata read a blank as a non-numeric character (a dot) or system missing" value, but others may read it as a zero. Analysts using statistical packages other than SAS or Stata should take care to distinguish between missing values and zeroes in programs used with these data because zeroes are often valid values on NSFG variables.

For many variables in the NSFG files, an *abridged* version of the complete universe statement is provided with the core routing information. These variables have nested routing statements, and for these variables, the most proximate routing statement will be described in the

universe statement. Since the universe statement contains the variable(s) that determined the routing into that question, users can trace back through the routing logic, that is, go to each preceding variable to see its routing statement and continue until the universe statement reads, "Applicable for all respondents."

For example: the question HC-9b ENDODIAG in the female questionnaire reads, "How many years ago were you first diagnosed with endometriosis?" It was asked of those who had ever been diagnosed with endometriosis. Thus, HC-9 ENDO("Endometriosis: Ever Diagnosed") is included in the universe statement for HC-9b ENDODIAG,

ENDODIAG	
LABEL	HC-9b How long ago was R diagnosed with endometriosis
VARIABLE TYPE	raw
UNIVERSE	Applicable if R reported diagnosis of endometriosis (HC-9 ENDO = 1)
LENGTH	8
VARNUM	52

ENDODIAG: Distribution	Value	n
Less than one year ago	1	32
1-4 years ago	2	57
5-9 years ago	3	76
10 years ago or longer	4	128
Don't Know	9	1
Inapplicable	.	5292
		5586

In addition to consulting the universe statement or "applicable specification" in the codebook documentation, you may also wish to consult the following resources available on the NSFG webpage:

- The CRQ (see section below entitled "**Description of Questionnaires**"), which contains more detailed specifications for the questionnaire:
  - For computed variables, the universe statements are drawn from the Flow Checks in which the variables are defined.
  - For the raw/asked variables as well as the computed variables, the questionnaires allow you to examine the sequencing and context of your variables of interest.
- The recode specifications, which are the source for the universe statements included in the codebook, and for the full details on how the recode was constructed and imputed.

Response Categories and Unweighted Frequencies: For categorical variables and several continuous variables in the NSFG, the codebook documentation lists all values, if there are any cases with those values in the data files, with descriptive value labels and unweighted frequencies (or counts of cases). For example, if no one responded "don't know" to a particular item, the "don't know" value will not be displayed in the codebook. To the extent possible, the exact wording of the questionnaire response choices is shown *(for example, see screen capture for AD-7b MARSTAT above).* Frequencies of variables that are not applicable for all respondents include the number of "inapplicable" cases. Most century month (date) and continuous variables have been collapsed *for display purposes* into more manageable groups, such as grouping

individual century months into ranges of years. The original values of these variables are intact in the file unless otherwise indicated in the variable labels. For example, the screenshot below for the variable cmfstuse shows that the reported century months have been collapsed into years for codebook display (Before 2017-2023), but the variable in the data file includes individual century month values as shown in the Value column.

CMFSTUSE	
LABEL	CM for date first used a method (if not at first sex) (Computed in FC E-2)
VARIABLE TYPE	computed
UNIVERSE	Applicable if first method use was after first sex (EB-2 FIRSTIME1 = 3,4,5, or 6 or EB-2 FIRSTIME2 = 3,4,5 or 6) or before first sex (EB-2 FIRSTIME2 = 1) or R refused or did not know the timing of her first method use (EB-2 FIRSTIME1 = RF/DK or EB-2 FIRSTIME2 = RF/DK)
NOTES	Use recode DATEUSE1
LENGTH	4
VARNUM	79

CMFSTUSE: Distribution	Value	n
Before 2017	865-1404	2294
2017	1405-1416	88
2018	1417-1428	88
2019	1429-1440	101
2020	1441-1452	76
2021	1453-1464	87
2022	1465-1476	88
2023	1477-1488	28
Refused	9998	33
Don't know	9999	51
Inapplicable	.	2652
		5586

Notes: For selected variables, the codebook entry will show a note with further information. The primary reasons for these notes are to indicate when there is a relevant recode that you should use *(see note pointing to the DATEUSE1 recode in the screen capture for cmfstuse above)*, or to indicate any disclosure risk reduction actions taken for this variable (described earlier in **"Protections to Minimize Risk of Disclosure for Individual-Level Data"**).

DESCRIPTION OF QUESTIONNAIRES

The 2022-2023 NSFG was conducted using CAPI for all respondents. As noted above, roughly 74% of respondents completed the entire survey independently on the web, and 26% were interviewed in person (i.e., FTF) by trained field staff. For those interviewed FTF, the final section of the interview (female section J and male section K) was self-administered using CASI. The NSFG webpage provides the full male and female questionnaires in two formats, with different levels of detail:

-- CAPI-Lite format
-- CAPI Reference Questionnaire (CRQ) format

CAPI-Lite Format

The male and female questionnaires are shown in their entirety, but with abridged representations of the question wording variants and shorter descriptions of skip patterns through the main survey. With this format, the emphasis is on getting a clear picture of how the questions were asked, in what order, and of which respondents, without showing every detail that was needed to program the questionnaire.

CAPI Reference Questionnaire (CRQ) Format

The CRQ shows all the detailed specifications that were used to program the NSFG questionnaires in Blaise Survey Software (www.blaise.com).

- All question wording variants are shown, along with the conditions defining when each variant was used. These include variants specific to FTF mode versus web mode, for some questions (designated with "FTFMODE=1" for FTF mode and "FTFMODE=2" for web mode).

- "Flow Checks" specify the precise routing through the survey based on earlier questionnaire items so that the appropriate next questions for the respondent appear onscreen. In addition, in some instances flow checks include the creation of new variables from one or more of the "raw" or "asked" variables. These are called "computed variables" and are described in other sections of the User's Guide (see **Description of Codebook, "Variable Type"**). The flow check specifies in detail how these computed variables were defined. A summary list of computed variables defined in each questionnaire section and those that are "passed forward" to be used for routing later in the survey can be found at the beginning of each section's CRQ.

- "Edit Checks," programmed into the instrument, attempt to catch and resolve data inconsistencies during the survey, rather than requiring resolution after data collection has ended. These consistency checks are generally located in the CRQ after the questions they are intended to reconcile. They are generally scripted for ease of use and enable the interviewer or web respondent to return to specific questionnaire items and correct them, if necessary. See also the User's Guide section on **Data Preparation for Public Use, "Logical Inconsistencies and Out-of-Range Values."**

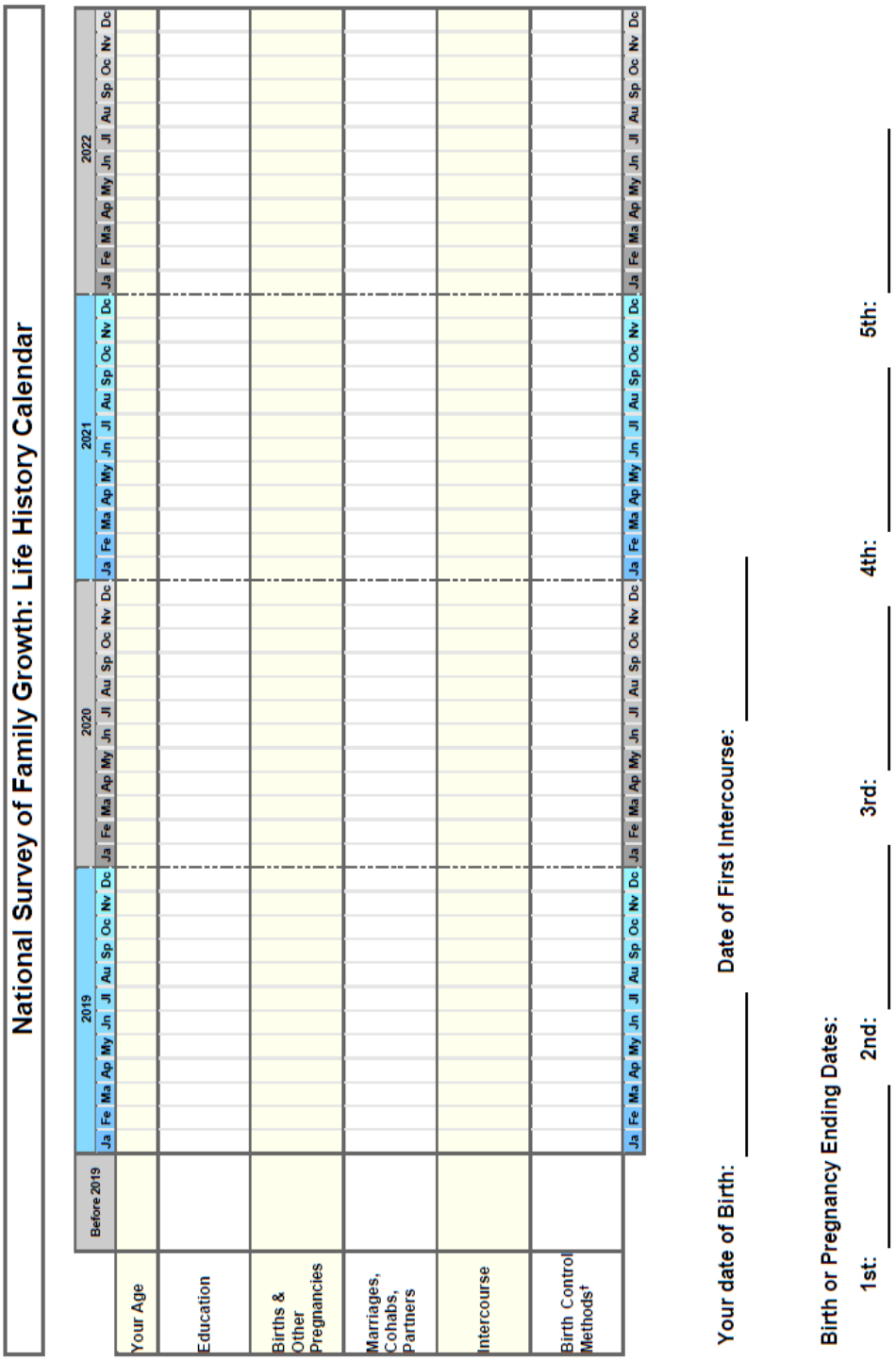
- Use of additional survey aids, such as Show Cards, Help Screens, and the Life History Calendar (female survey only), and onscreen instructions for interviewers are noted on individual questionnaire items.

    - Help Screens: If a question-specific help screen was available for an item, the CRQ indicates "[HELP AVAILABLE]." This was displayed on the screen in the instrument for FTF mode and for web mode was displayed as "?" on the screen and hyperlinked to the help text.
    - Show Cards: Relevant for FTF mode, if the item's response choices were to be

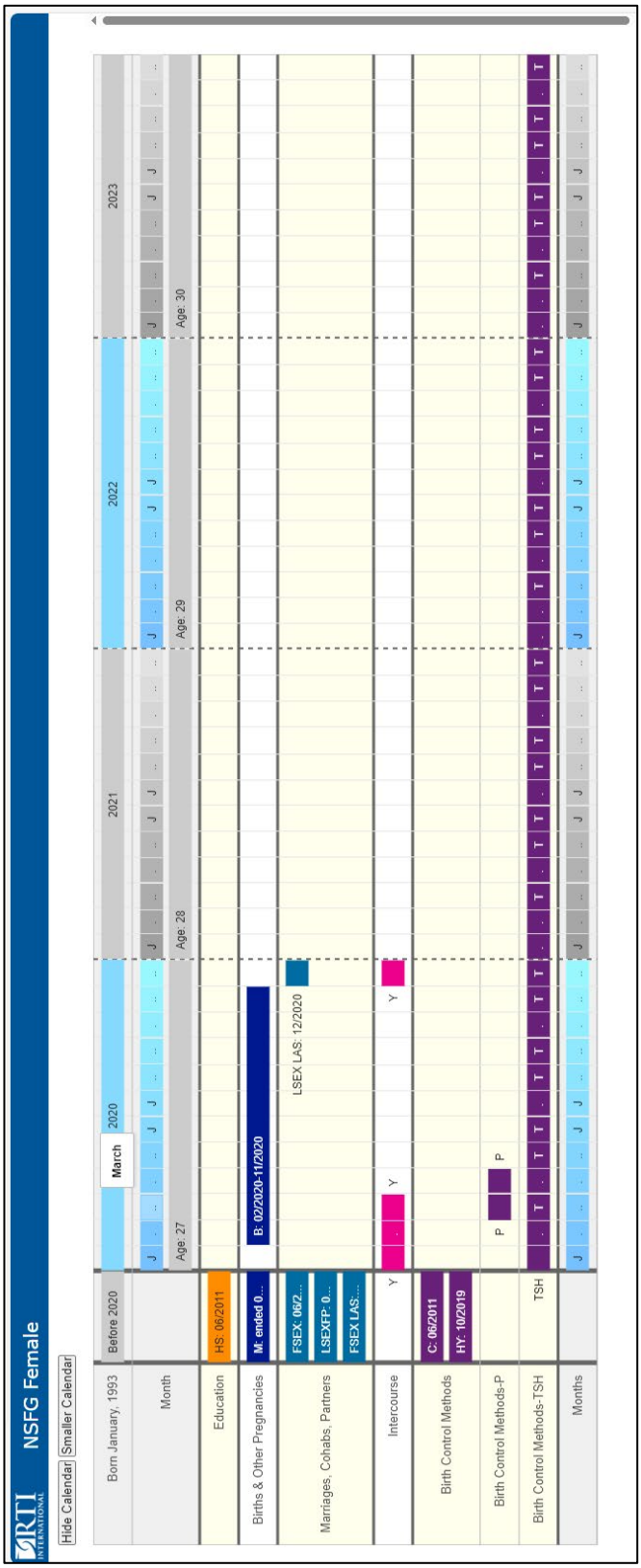
shown on a Show Card in the interviewer's show card booklet, the CRQ indicates the number of the show card along with the response categories. Response categories were presented on the screen for web respondents similar to how they are shown in the CRQ.

- Interviewer Instructions: Also shown are the onscreen instructions for interviewers that accompanied many of the questions. These instructions are preceded by "IF FTFMODE=1..." indicating the instruction would show only for FTF mode.
- Life History Calendar (used only with female respondents): On the next 2 pages, there is an example of the paper Life History Calendar used in FTF interviews and an example of the electronic version shown on screens in web surveys. These versions reflect calendars used for interviews or surveys completed in 2022. Since the female questionnaire includes a greater number of questions about dates and the relative timing of events than the male questionnaire, the Life History Calendar was only used for female respondents. The Life History Calendar has been shown to improve recall of dates by anchoring responses to key life events for the respondent.

Paper Life History Calendar used with Face-to-Face Surveys with Female Respondents



Electronic Life History Calendar used with Web/Online Surveys for Female Respondents: Example Screen



ACKNOWLEDGMENTS

As noted in the survey background section above, the NSFG has been designed, administered, and disseminated by the National Center for Health Statistics since 1973, in collaboration with several other agencies of the U.S. Department of Health and Human Services (DHHS). (NCHS became part of CDC in 1987.) The 2022-2023 NSFG was jointly planned and funded by the following agencies within the DHHS:

- CDC/National Center for Health Statistics (NCHS)
- Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD)
- Office of Population Affairs (HHS/OPA)
- Office on Women's Health (HHS/OWH)
- Children's Bureau of the Administration for Children and Families (ACF/CB)
- Office of Planning, Research, and Evaluation within ACF (ACF/OPRE)
- CDC/NCHHSTP/Division of HIV Prevention (DHP)
- CDC/NCHHSTP/Division of STD Prevention (DSTDP)
- CDC/NCHHSTP/Division of Adolescent and School Health (DASH)
- CDC/NCCDPHP/Division of Cancer Prevention and Control (DCPC)
- CDC/NCCDPHP/Division of Reproductive Health (DRH)
- CDC/NCIPC/Division of Violence Prevention (DVP)

The NSFG team at NCHS holds primary responsibility for all aspects of the survey design, public-use data and documentation preparation, and data dissemination, including published reports and key statistics on the NSFG webpage. The NCHS NSFG team works closely with the contractor RTI International on the sample design, data collection, and data processing and documentation for the survey (Contract #200-2020-09732).