Public-Use Data File Documentation

2017-2019
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
October 2020

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(d)] and the Confidential Information Protection and Statistical Efficiency Act or CIPSEA [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 slightly modified through statistical perturbation. These modifications are intended to prevent definitive identification of individual respondents. 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.

TABLE OF CONTENTS

APPENDICES (provided in separate PDF files posted on NSFG webpage)

GENERAL INFORMATION FOR USERS OF THE 2017-2019 NSFG PUBLIC-USE DATA

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

- An overview of the 2017-2019 NSFG and brief background on previous NSFG surveys
- How to access NSFG data and documentation files
- How the NSFG data files are organized
- Information on sampling weights and variance estimation
- How the NSFG interview data were prepared for public use
- Descriptions of the codebooks and questionnaires

The User's Guide also contains seven appendices providing essential technical documentation and reference information:

1. File indexes (i.e., listings of each variable in the order in which they appear on the public-use file, along with a brief description)
2. Topic-specific notes for 2017-2019
3. Recode specifications for specially constructed and imputed variables
4. Recode crosswalks by sex and across recent NSFG survey years
5. Summary of questionnaire changes since 2015-2017
6. Frequently Asked Questions (FAQ)
7. Listing and description of analytic variables suppressed, modified, or created for public use.

As noted above, **Appendix 6** of this User's Guide provides a list of **frequently asked questions (FAQ)** about the NSFG, geared toward current or prospective users of the data file. Here are the primary highlights from the FAQ:

- **Use sample weight (e.g., WGT2017_2019) and design variables (SECU & SEST)** 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.** They have been edited carefully, and missing data have been imputed. Some of the most commonly used recodes are listed on page 19.

- In addition to this 2017-2019 NSFG User's Guide with its seven appendices, the **NSFG website** also provides the following materials:
    - the **codebooks** (online codebook called "Webdoc")
    - the **questionnaires** in two levels of detail
    - downloadable **ASCII data files**

- **program statements** to read in the ASCII data files using SAS, Stata, and SPSS
- **variance estimation examples** using 2017-2019 data that can be adapted for your own research purposes
- **case weights** for analyzing combined 4-year, 6-year, and 8-year data files (2011-2015, 2013-2017, 2015-2019, 2011-2017, 2013-2019, and 2011-2019)

- Before embarking on analyses based on combined NSFG data for all or part of 2011-2019, consult the **"2011-2019 Combined Files: Selected Data and Documentation"** webpage for specific guidance and relevant case weights.

- If you cannot find an answer to your question in the NSFG User's Guide, codebooks, questionnaires, or webpage, please contact NSFG staff at nsfg@cdc.gov.

- The 2017-2019 NSFG sample design, continuous fieldwork plan, and other procedures of the survey are similar to what was used in 2006-2010, 2011-2013, 2013-2015, and 2015-2017. Until further information is published on the NSFG webpage specific to the survey design and operations of the 2017-2019 NSFG, refer to the section below on **"Sample Weights and Variance Estimation,"** the methodology reports available on the webpage for 2015-2017, 2013-2015, and 2011-2013, as well as the Series 1 and Series 2 technical reports from the 2006-2010 data release available in PDF format on the NSFG website.

BACKGROUND AND OVERVIEW OF THE 2017-2019 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, as well as key personnel at NCHS and at the contractor organization for this file release, the University of Michigan's Institute for Social Research.)

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; and
- attitudes about sex, childbearing, and marriage.

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

Cycle	Year	Scope	Number of Interviews	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	Blacks, Hispanics, 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	Blacks, Hispanics, 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	Blacks Hispanics 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	Blacks Hispanics Teens	W=80 min M=60 min	$40	63.4% W=65.2% M=61.4%

The data included in this 2017-2019 NSFG public-use file release are based on a multi-stage probability-based, nationally representative sample of the household population aged 15-49. Fieldwork for the 2017-2019 NSFG was conducted from September 2017 through September 2019, based on survey protocol and informed consent procedures approved by the NCHS Research Ethics Review Board (protocol #2015-12). One sample respondent per household was selected based on screening interviews in NSFG sample households. In-person interviews were conducted with 6,141 women and 5,206 men 15-49 years of age for a total sample size of 11,347 over the 2-year fieldwork period. All interviews were conducted by female interviewers trained specifically for the NSFG survey using laptop computers programmed with the survey questionnaires, an approach known as computer-assisted personal interviewing (CAPI). Informed consent-related materials for the 2017-2019 NSFG are posted on the NSFG website. Electronically signed parental permission and minor assent were obtained for all minor respondents aged 15-17. Adult respondents in 2017-2019 could provide their consent without signature. In 2017-2019, the interviews for female respondents averaged 76.4 minutes in length, and the interviews for male respondents averaged 51.9 minutes, both within the limits of 80 minutes for females and 60 minutes for males approved by the Office of Management and Budget (NSFG OMB No. 0920-0314).

Response Rates: The overall response rate for 2017-2019 NSFG for ages 15-49 was 63.4%; 65.2% for women and 61.4% for men. The response rate for female teenagers ages 15-19 was 66.0% and 65.3% for male teenagers. Response rates for ages 15-44 for 2017-2019 were 63.7% overall; 65.5% for women and 61.7% for men. The calculation of response rates for the NSFG under the continuous fieldwork design that began with the 2006-2010 NSFG, has been similar for each data release.

Specific details on how the 2017-2019 survey was designed and conducted are available on the NSFG webpage (expected by December 2020) in the section **"Design and Data Collection Methods."** The titles of the four summary reports on design and data collection to be released for the 2017-2019 NSFG are listed below. The principles of continuous interviewing and many ongoing design features have remained the same since 2006.

- 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): Sample Error Estimation Design
- 2017-2019 National Survey of Family Growth (NSFG): Weighting Design Documentation

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

The public-use data and documentation for the 2017-2019 NSFG are available on the NSFG website. Documentation includes this User's Guide, the questionnaires, and a link to the interactive, online codebook documentation ("Webdoc"), which shows separate entries for each variable on the files.

The public-use data are contained in three data files:

- **Female respondent file** (one record or observation per interviewed female)
- **Female pregnancy (interval) file** (one record per pregnancy of interviewed females)
- **Male respondent file** (one record per interviewed male)

New for the 2017-2019 NSFG:

- There are a number of new variable suppressions and modifications related to disclosure risk reduction for the 2017-2019 public-use files. See the section further below on **"Protections to Minimize Risk of Disclosure of Individual-Level Data"** as well as **Appendix 7** of this User's Guide for more information.

ORGANIZATION OF THE 2017-2019 NSFG PUBLIC-USE DATA FILES

The public-use data for the 2017-2019 NSFG are provided as three separate ASCII files.

FILE CHARACTERISTICS	**Number of Records (observations)**	**Record Length (number of columns)**	**Number of Variables**
Female respondent file File = 2017_2019_FemRespData.dat (one record per female respondent)	6,141	3,839	2,609
Female pregnancy (interval) file File = 2017_2019_FemPregData.dat (one record per pregnancy reported by female respondents)	10,215	251	171
Male respondent file File = 2017_2019_MaleData.dat (one record per male respondent)	5,206	4,088	3,009

The **Female Respondent file** contains one record for each of the 6,141 women aged 15-49 interviewed in the survey.

The **Female Pregnancy (Interval) file** contains one record for each of the 10,215 pregnancies reported by female respondents. Pregnancy records are based on both completed pregnancies (those that have reached an outcome such as live birth, stillbirth, ectopic, miscarriage, or induced abortion) and current pregnancies (ongoing at time of interview). Each pregnancy record contains information about the characteristics of that pregnancy and method use and wantedness before that pregnancy. That is, in the Female Respondent file the unit of analysis is the female respondent, and in the Pregnancy file the unit of analysis is the pregnancy or pregnancy interval.

The **Male Respondent file**, often just referred to as the male file, contains one record for each of the 5,206 men aged 15-49 who were interviewed; the male respondent is the unit of analysis.

Program statements are provided on the NSFG webpage to read the ASCII data into SAS, SPSS, and Stata, and to apply appropriate variable labels and formats that include value labels. The formats provided within the program statements are for user convenience or ease of display only. These formats do not always reflect the actual values in the public-use dataset and sometimes condense variable values into groups Data users should make their own decisions about whether to use the formats provided in the program statements.

Data Layout for Each File

The following is a listing of the major sections in the three public-use files from the 2017-2019 NSFG. Not all items asked in the NSFG questionnaires could be included on the public-use files due to disclosure risk concerns. **Appendix 1** (File Indexes) provides more detail, including short descriptions 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. **Appendix 7** provides lists of all analytic variables available only through the NCHS Research Data Center, as well as descriptions of all variables modified or created for public use.

FEMALE RESPONDENT FILE *– information for each female respondent interviewed*

- Respondent ID (CASEID) and selected screener variables
- Questionnaire Data (including computed variables) for Sections A-J
    A. Background and demographic information
    B. Pregnancy and adoption-related information
    C. Marital and relationship history; first sexual intercourse; recent partners; sex education
    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. Infertility services and reproductive health
    I. More background information: demographic information, access to health care & attitude questions
    J. Audio CASI: general health measures; pregnancy re-reporting; cigarette, alcohol & other drug use; STD/HIV-risk behaviors; sexual orientation & attraction; income and economic insecurity
- Recodes (constructed, imputed variables) & imputation flags *(including key recodes describing pregnancies, provided for user convenience)*
- Weights & related variables *(including sample design variables)*
- Century month of interview and related variables

FEMALE PREGNANCY (INTERVAL) FILE *– information for each pregnancy reported by female respondents (including current)*

- Respondent ID (CASEID)
- Pregnancy Order (PREGORDR)
- Questionnaire Data (including computed variables) for Sections B & E
    - B: pregnancy outcomes and years, prenatal care, low birth weight, sources of payment for delivery, breast-feeding
    - E: contraceptive use in the pregnancy interval and wantedness of the pregnancy
- Recodes (constructed, imputed variables) based on pregnancy-specific variables & associated imputation flags *(along with key recodes and other variables from respondent file, provided for user convenience)*

- Weights & related variables *(including sample design variables)*
- Century month of interview and related variables

MALE RESPONDENT FILE *– Information for each male respondent*

- Respondent ID (CASEID) and selected screener variables
- Questionnaire Data (including computed variables) for Sections A-K
    A. Background and demographic information
    B. Ever sex, sex communication and education, vasectomy and physical ability to father children, number of sexual partners, enumeration and relationship with up to 3 recent (or last) sexual partner(s)
    C. Current wife or cohabiting partner: years of key dates of marriage or cohabitation; contraceptive use with her; children
    D. Recent (or last) sexual partner(s) (up to three): years of key relationship dates, contraceptive use with her, children, 1st sexual partner ever
    E. Former wives and first premarital cohabiting partner: years of key relationship dates, children
    F. Other biological children, other adopted children, other pregnancies
    G. Fathering: Activities with the youngest child he lives with and the youngest child he lives apart from
    H. Desires and intentions for future biological children
    I. Health conditions, access to health care, and receipt of health services
    J. More background information: demographic information & attitude questions
    K. Audio CASI: pregnancy reporting; cigarette, alcohol & other drug use; STD/HIV-risk behaviors; sexual orientation & attraction; income and economic insecurity
- Recodes (constructed, imputed variables) & imputation flags
- Weights & related variables *(including sample design variables)*
- Century month of interview and related variables

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. The documentation referenced earlier on page 8, **"Design and Data Collection Methods,"** provides more detailed information about the 2017-2019 sample weights. Given that the sample design for the 2017-2019 NSFG is largely the same as that for the 2006-2010, 2011-2013, 2013-2015, and 2015-2017 files, much of the design information published for those earlier file releases is also applicable for 2017-2019.

Sampling Weights

Each respondent in the 2017-2019 NSFG sample represents a different number of people in the U.S. household population, and this number is indicated in the respondent's sampling

weight. There are several factors that lead to variation in the size of the weights. For example, Hispanic persons, 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. Sampling weights adjust for these unequal probabilities of selection for different population subgroups. The sampling 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 adjusted to U.S. Census Bureau estimates of the number of persons in age-sex-race-ethnicity subgroups. Data users should use the weights in all analyses to obtain accurate estimates. Using the weights will permit replication of the nationally representative estimates that appear in published NCHS reports.

Each of the 2017-2019 data files have a weight variable called "WGT2017_2019" with values for each of the 6,141 female and 5,206 male respondents who completed NSFG interviews in 2017-2019. When correctly applied for the full set of cases, this "WGT2017_2019" variable yields estimates representative of the 72.7 million women and 72.2 million men in the household population aged 15-49 of the United States in each dataset at the approximate midpoint of 2017-2019 interviewing (July 2018). The weighted population sizes will of course be smaller if the user conducts analyses for other population subgroups.

For analyses where larger sample sizes are needed to achieve sufficient statistical reliability, users may wish to combine data from more than one NSFG file release. For further guidance, as well as appropriate case weights, for combining NSFG data for 2011-2019, please consult the **"2011-2019 Combined Files: Selected Data and Documentation"** webpage.

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=WGT2017_2019/1000

In addition to using the sampling weight variable WGT2017_2019, researchers must use the design variables for the sampling stratum (SEST) and cluster (SECU) to obtain correct standard errors for their estimates. These values of SEST and SECU would be used regardless of whether the user is analyzing a two-year files (with the two-year file weight) or combined files for 2011-2019 with the appropriate combined-file weight, as described and available on the **"2011-2019 Combined Files: Selected Data and Documentation"** webpage.

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 2017-2019 NSFG, the sampling variance measures variation from the full population-based parameters due to interviewing the NSFG sample of 11,347 respondents instead of all 144 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 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 to terminate or lead to other errors. Three variance estimation examples using SAS and Stata for the 2017-2019 file are posted on the NSFG webpage. Example 3 shows how to create a subpopulation variable in SAS and Stata.

OVERVIEW OF DATA QUALITY IN THE NSFG

As measured by amounts of missing data and inconsistent data, data quality in the 2017-2019 NSFG is high, as it was in previous data years. This high quality was obtained through:

-- Questionnaire design work, including careful specification, testing, and incorporation of lessons learned from the past NSFG data collection periods;
-- Consistency checks built into the interview that allowed potential data problems to be resolved in the field rather than after data collection;
-- Evaluation of monthly 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 interview data for public use. Some of these actions were taken simply to make the data more 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 fieldwork, logical consistency across data items was maintained through "edit checks" built into the programs that ran the male and female questionnaires. These edit checks alerted the interviewer to inconsistent or out-of-range entries and required that she attempt to correct the entry, usually by working with the respondent. Out-of-range values are minimal in the 2017-2019 NSFG (as in past files) because valid ranges are specified and programmed into the instrument to the extent possible, and values outside that range are rejected or signaled by the computer.

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 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 his or her responses in case they were given in error. If the respondent says that the information is accurate, the interviewer can override or suppress the inconsistency warning box and enter a brief comment to explain the situation. In rare cases, the interviewer herself may have misunderstood the edit check and mistakenly suppressed it. 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 interviews, and as a result, some inconsistencies in the data could not be eliminated.

In addition to edit checks, other aspects of the questionnaire designed to maximize consistency *during* data collection were: 1) "summary screens" before or after key sections, reminding the respondent of events and dates reported earlier, and 2) life history calendars provided to female respondents as a visual aid for recording and remembering the chronology of events.

As in prior NSFG file releases, the process of checking for consistency within the 2017-2019 data 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 by the interviewer to indicate that the respondent could not or would not provide an answer to a question. "Not ascertained" refers to rare instances in which a question was erroneously skipped during the interview. 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 ACASI (in Section I for females and in Section J for males). The small number of respondents who did not complete the ACASI section, partially or completely, will have "not ascertained" values assigned to all variables after their break-off point.

Depending on the column length of the original data items:

- "don't know" values are coded 9, 99, 999, 9999, or 99999
- "refusal" 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 **Appendix 3** (Recode Specifications)).

Century-Month Coding for Dates

During the interview, dates of events were collected as month and year. For every date asked in the interview, 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 2006 is (106 x 12) + 7 = 1279.

The century month form is convenient for computing intervals between dates, 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. The following codes were used for the 3 types of missing data on century-month date variables:

9997 = Not ascertained
9998 = Refused
9999 = Don't know

When a season was reported on any month variable in the NSFG interview, the months shown below were consistently assigned to enable the construction of a century month value and facilitate subsequent routing through the questionnaire.

Winter = 1 (January)
Spring = 4 (April)
Summer = 7 (July)
Fall = 10 (October)

If a respondent said "don't know" or refuses (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 805 (January 1967) through 1440 (December 2019) are shown in the table below with the years from 1967 through 2019 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 interviews for the 2017-2019 NSFG were conducted between September 2017 (century month 1413) and September 2019 (century month 1437).

In response to disclosure risk concerns associated with a number of dates collected in the NSFG interview, as first done in 2015-2017, the public-use files for 2017-2019 no longer include all century-month dates collected in the interview 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
1967	805	806	807	808	809	810	811	812	813	814	815	816
1968	817	818	819	820	821	822	823	824	825	826	827	828
1969	829	830	831	832	833	834	835	836	837	838	839	840
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

Other-Specify Coding

In the 2017-2019 NSFG, as in past surveys, a small number of questions contained items to which respondents could specify a response other than those provided, and this response was typed in verbatim by the interviewer. In most cases, these questions appear in the questionnaires with a "_SP" at the end of their variable name or a "SP_" at the beginning, and the question text and response type distinguish these as open-ended questions. These verbatim response variables are NOT included on the public-use files because of confidentiality concerns, but the essential information for data users has been coded in a manner that does not pose risk of disclosure. In all cases, responses that were clearly codable using a pre-specified category (one of the categories offered in the questionnaire and shown to respondents on a show card) were edited (or "back-coded") into that pre-existing category. In a few instances, the responses were categorized, and a new variable was created that contained only these additional responses.

In summary, the actual verbatim responses for other-specify variables are not included in the public-use data files, but they have all been reflected in some way in the associated, closed-ended numeric variables that are included on the public-use file. Questions where new categories or new variables were created based on other-specify responses have a short description in a "Note" entry on their codebook page (see also **Description of Codebooks** further below). Please consult the questionnaires for 2017-2019 as posted on the NSFG webpage for further details on these questions with "other (specify)" options. **Appendix 2** of the User's Guide also includes specific information related to some of these items.

Recodes and Imputation

*(also see **Appendix 1** (File Indexes), **Appendix 3** (Recode Specifications), and **Appendix 4** (Recode Crosswalks))*

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. (Recodes comprise about 10% of the variables in these files.)

Some recodes are simple, while others are complex. Some recodes may simply be transferred from single questionnaire items and imputed if missing (for example, RCURPREG, whether respondent is currently pregnant). Other recodes are based on multiple questionnaire items and may involve more intricate logic to define (for example, CONSTAT1, the respondent's current contraceptive status).

Before using the original variables or constructing their own summary variables, analysts are encouraged to **check to see if a relevant recode exists**. Many of the raw or computed variables that have a recode corresponding to them will have a **note on their Webdoc codebook pages** 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
RMARITAL	Informal marital status
EDUCAT	Education (number of years of schooling)
HIEDUC	Highest completed year of school or highest degree received
HISPANIC	Hispanic origin, regardless of race
RACE	Race of respondent, regardless of 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	Poverty level income
TOTINCR	Total income of R's family
PUBASSIS	Whether R received public assistance in the calendar year before the interview
HADSEX	Whether R has ever had sexual intercourse with opposite sex
VRY1STAG	Age at first intercourse
VRY1STSX	Date (century month) of first intercourse
CONSTAT1	Current contraceptive status (females only)

Besides the above list, other sources to check are the File Indexes in **Appendix 1** or the Recode Specifications in **Appendix 3** to see if a relevant recode exists. If you want to see whether there are comparable recodes between males and females or across NSFG data years, **Appendix 4** contains 3 crosswalks for this purpose:

- Appendix 4a: recodes for males and females in 2017-2019 NSFG (arranged by topics)
- Appendix 4b: recodes for females across 2011-2013, 2013-2015, 2015-2017, and 2017-2019 NSFG (arranged by section of questionnaire)
- Appendix 4c: recodes for males across 2011-2013, 2013-2015, 2015-2017, and 2017-2019 NSFG (arranged by section of questionnaire)

The frequency of missing values for the recoded variables in 2017-2019 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 imputed.

Most missing recode values were assigned using **regression** imputation software in which multiple regression is used to predict a value for the case using other variables in the data

set as predictors. Regression imputation follows the same logical constraints that are built into the original 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 2017-2019 NSFG was similar to that used in prior releases; the **"Summary of Design and Data Collection Methods"** report on the NSFG webpage 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.

Finding recodes in the data file and codebook: As shown in **Appendix 1** (File Indexes), the recodes and their imputation flags are clustered together near the end of each of the three data files. Recodes can be distinguished in the codebook documentation by the "Variable type" displayed on the codebook page between the variable name and the variable description. The word "recode" also appears at the end of the variable's "question text" or short description.

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 fieldwork 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 2017-2019 were submitted to the NCHS Disclosure Review Board (DRB). In brief, the disclosure risk protections taken for the 2017-2019 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 <u>not</u> 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).

- REGION of residence at time of interview, is available only 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, military service, 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.
    - Since the public-use file includes only years of pregnancy conceptions and outcomes, with the century month dates available only through the NCHS RDC, four types of inter-pregnancy interval variables, in categorical form, have been created for public use.

- **In keeping with the changes made in 2015-2017, the 2017-2019 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 2017-2019, and the full inventory (with further details) of all variables suppressed, modified, or created for public use in 2017-2019 can be found in Appendix 7. Appendix 7 also lists all restricted-use analytic variables that are available only through the RDC. (Also see Appendix 1 where all variables with disclosure risk reduction (DRR) actions have been asterisked, and those with new DRR actions in 2017-2019 have been highlighted in yellow):

    - The non-voluntary sexual intercourse series in ACASI (female JE and male KF and KI) have been suppressed for public use.

- The level of pregnancy-specific detail included on the female pregnancy file for public use has been significantly reduced – in particular, sex of liveborn children has been suppressed and gestational length has 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, age at first sexual intercourse has been bottom-coded, ages of nonbiological children reported by men and women have been made categorical, ages at selected preventive health services have been bottom-coded).

Whenever variables have been suppressed, modified, or newly created for reasons of disclosure risk reduction, a note has been included in the online codebook, Webdoc (described in the next section). 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. See Appendix 7 of the User's Guide for further information."

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. See Appendix 7 of the User's Guide for further information."

> "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. See Appendix 7 of the User's Guide for further information."

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. See Appendix 7 of the User's Guide for further information."

In addition to these codebook notes, the file indexes provided in **Appendix 1** include an asterisk when a disclosure risk reduction action was taken for that variable. These asterisks are highlighted in yellow if the disclosure risk reduction action is new for 2017-2019.

As a final step to prevent identification of individual respondents, 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 list of these restricted-use, analytic variables is provided in **Appendices 7a and 7b**, and further details on all variables that have been modified or created for public use are provided in **Appendix 7c**. 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.

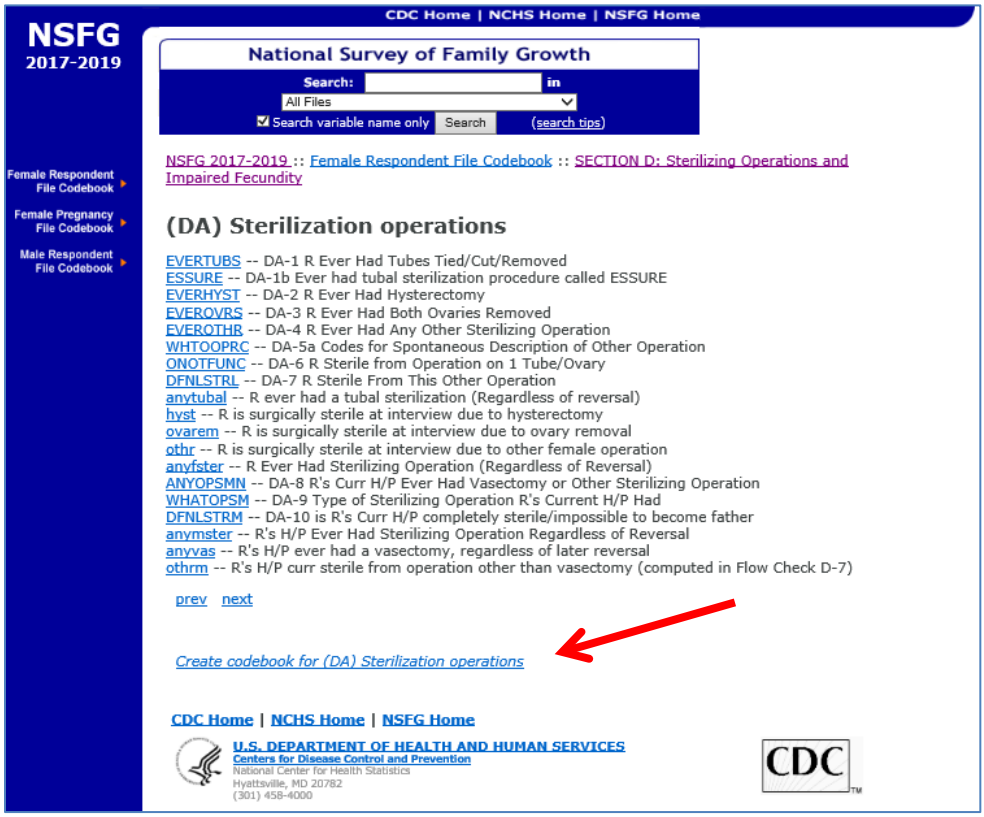
DESCRIPTION OF CODEBOOKS

Overview

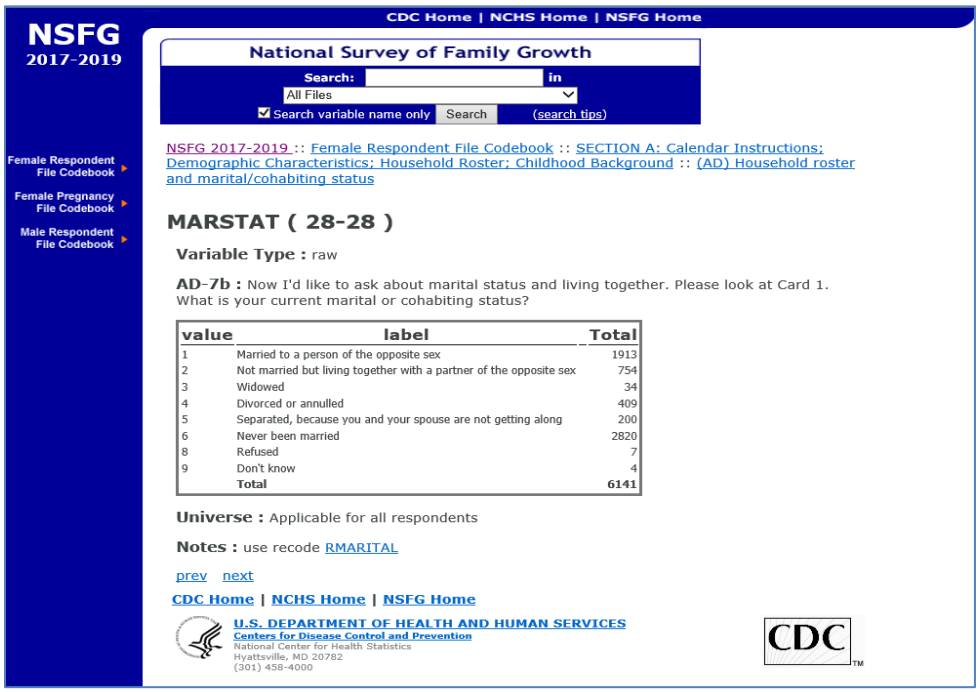
Codebooks for the NSFG provide essential information for each variable included in the public-use files. The elements of the codebook, such as variable type and universe statements, are described further below.

Webdoc is a web-based tool that permits online, interactive access to the NSFG codebooks and allows easy access to all variables, quick navigation between different sections of the instrument (through hyperlinks and through menu-like lists of data files, sections and series), and searching for specific words or phrases or for specific variables. For recoded variables, direct links are provided to the recode specifications. These recode specifications are the same as those provided in **Appendix 3** on the NSFG webpage.

To generate an excerpt for specific sections or subsections of the codebook use the hyperlink at the bottom of the screen that says "Create codebook for <Section/Subsection>." The resulting file can be saved in PDF, printed, or simply viewed on the screen. For example, as shown below, at the bottom of the Webdoc page listing the variables in sub-section, "Sterilization Operations (DA)" if you click on *Create codebook for (DA) Sterilization Operations*, a screen with just the variables in the DA series from Female Section D is created.



Below is an example page from Webdoc displaying the detailed codebook information for the raw variable (and question) AD-7b MARSTAT. The specific elements of the codebook are described further below.



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
- Column locations
- Question text
- Universe statement
- Response categories and unweighted frequencies
- Link to recode specifications, where applicable
- 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). *(For example, in the Webdoc screen capture above, the variable name and question name are the same – MARSTAT.)* Recode and intermediate variable names correspond to those found in the recode specifications (**Appendix 3**). 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, BD-8 PAYBIRTH in the CRQ for Female Section B 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 2017-2019 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 and question text also.

Variable Type:

On the codebook page for each variable, underneath the variable name, is "Variable Type." There are five basic variable types included in the NSFG files -- "raw," "computed," "recode," "intermediate," and "computed in post-processing":

1. A **raw** variable refers to a question that was asked during the interview (the majority of variables in the data files are raw). *(For example, in the Webdoc screen capture above, AD-7b MARSTAT is labeled as a raw variable.)*
2. A **computed** variable is a variable computed as part of the Blaise-programmed survey instrument during the interview, 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 variables, and has missing values imputed.
4. An **intermediate** variable is one defined in the specifications for certain recodes. These are sometimes included in the public-use data files because they can be useful for analysts.

Intermediate variables are defined in the process of constructing a recode. They are very few in number.

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.

Column Locations:

Next to each variable name, the codebook page gives the column locations in parentheses. *(For example, in the Webdoc screen capture above, the column location for MARSTAT is 28 on the female respondent file.)*

Question Text:

Question text is either the actual question wording for a raw, asked question in the interview or the short description of all other types of variables in the file. The wording of the survey question is shown in the codebook for "raw" variables and is preceded by the question number. Any question wording variants are presented, sometimes in collapsed form. For computed variables (computed as part of the Blaise-programmed survey instrument), the "question text" includes the "Flow Check" number from the CAPI Reference Questionnaire (CRQ) where the computed variable was defined. For example, "(Computed in Flow Check E-13b)" indicates that the variable was defined in Section E, Flow Check E-13b. CRQs can be consulted to see these flow checks and the specifications for defining each computed variable. For recodes, intermediate variables, and variables computed in post-processing (all variables not represented in the questionnaires), the question text corresponds to the variable's short description from the Recode Specifications (see **Appendix 3**).

When variables are part of an array or loop, the question text on the codebook page indicates what is being referenced, just as the variable descriptions do in the File Indexes (**Appendix 1**). For example, the question text for male CG-5 CWPCHSEX2 makes clear that the variable applies to the second biological child the respondent had with his current wife or cohabiting partner.

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 Webdoc screen capture for AD-7b MARSTAT above.)*

Cases with inapplicable values on any variable are coded as blank or "system missing." 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 EA-12 ECTIMESX in the female questionnaire reads, "How many different times have you used emergency contraception?" It was asked of those who had ever used emergency contraception. Thus, EA-11 MORNPILL ("whether R ever used emergency contraception") is included in the universe statement for EA-12 ECTIMESX, and hyperlinked. Clicking on this variable takes you to the codebook page for EA-11 MORNPILL. MORNPILL was only asked of those who had ever had sex, so its universe statement contains the computed variable for ever had sex: "rhadsex." Clicking on this takes you to the page for this computed variable, which is "applicable for all respondents."

All public-use file variables referenced in the universe statements are hyperlinked in Webdoc (see description above) so that users can go directly to their codebook pages. The variable names will not be hyperlinked in any PDF file generated from Webdoc, but users should still find it straightforward to find the relevant codebook entries for variables referenced in the universe statement. To make it easier to locate these variables, recall that question numbers precede the names of all raw variables. Also, the names of computed variables appear in lower case, and the names of raw variables and recodes appear in upper case. If a universe statement references variables that are not included on the public-use file those variables are not hyperlinked.

In addition to consulting the universe statement or "applicable specification" in the codebook documentation, you may also wish to consult:

- The CAPI Reference Questionnaire (CRQ; see section below entitled "**Description of Questionnaires**"), which contains more detailed specifications for the questionnaire. The universe statements for the computed variables 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 (**Appendix 3**), 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 Webdoc 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. Response categories are not displayed unless at least 1 case reported such a response, and this also applies to "refused" (8, 98, etc.), "don't know" (9, 99, etc.), and "not ascertained" (7, 97, etc.) responses.

Link to recode specifications:

For every recode variable, there is a direct link in Webdoc to the specification for that recode in **Appendix 3**. These hyperlinks will not be active in any generated PDF files for the codebooks.

Notes:

For selected variables, the codebook page 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 RMARITAL recode in the Webdoc screen capture for AD-7b MARSTAT above),* to describe special circumstances related to the response categories (for example, see female HE-4 PLCHIV), 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 NSFG webpage provides the male and female questionnaires in two formats, with different levels of detail:

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

Both formats represent the basic content and routing of the full NSFG interviews, including the computer-assisted personal interviews (CAPI) administered by interviewers and the audio computer-assisted self-interviews (ACASI) that respondents completed on their own. However, each format of the questionnaire offers users a different level of detail on how the interview was conducted.

CAPI-Lite Format

The male and female interviews are shown in their entirety, but with abridged representations of the question wording variants and shorter descriptions of skip patterns through the interview. 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 (see http://www.blaise.com/). While the entire female and male CAPI-Lites are contained in two PDF documents, the considerably longer CRQs are provided on the NSFG webpage as separate PDF files corresponding to each section of the questionnaire.

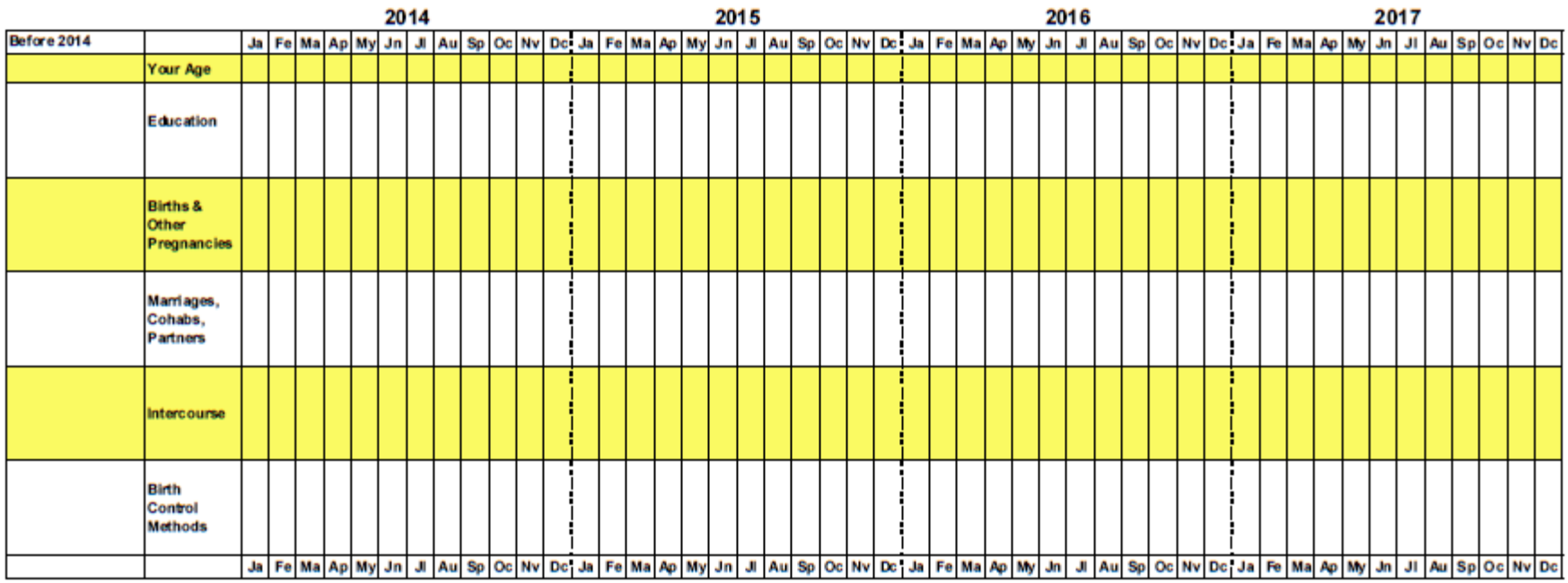
- All question wording variants are shown, along with the conditions defining when each variant should be used.

- "Flow Checks" specify the precise routing through the interview 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 can be found at the beginning of each section's CRQ, and those that are "passed forward" to be used for routing later in the interview are listed at the end of each section's CRQ.

- "Edit Checks," programmed into the instrument, attempt to catch and resolve data inconsistencies during the interview, 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 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 interview only), are noted on individual questionnaire items. For example, if a question-specific help screen was available for an item, the CRQ indicates "[HELP AVAILABLE]." 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. Also shown are the onscreen instructions for interviewers that accompanied many of the questions. An example of the Life History Calendar is shown on the next page. The version shown is for interviews conducted in 2017. Since the female interview includes a greater number of questions about dates and the relative timing of events than the male interview, the Life History Calendar was only used in the female interview. The Life History Calendar has been shown to improve recall of dates by anchoring responses to key life events for the respondent.

National Survey of Family Growth: Life History Calendar



Your Date of Birth: _____

Date of First Intercourse: _____

Birth or Pregnancy Ending Dates:

1st: _____ 2nd: _____ 3rd: _____ 4th: _____ 5th: _____

USER SUPPORT

Most commonly asked questions about the NSFG are addressed in the "Frequently Asked Questions" included in this User's Guide in **Appendix 6** and throughout the main text of this User's Guide. If, however, you have reviewed this Guide and its appendices thoroughly and you still have a question, please contact the NSFG team at nsfg@cdc.gov.

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 2017-2019 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/AIDS Prevention (DHAP)
- 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/NCCDPHP/Division of Nutrition, Physical Activity and Obesity (DNPAO)
- CDC/National Center for Birth Defects and Development Disabilities (NCBDDD)

The NSFG team at NCHS holds primary responsibility for all aspects of the survey design, public-use file and documentation preparation, and data dissemination as well as published reports and key statistics on the NSFG webpage. The NCHS NSFG team works closely with the contractor on the sample design and data collection for the survey. Fieldwork for the 2017-2019 NSFG was conducted under contract with the University of Michigan's Institute for Social Research (ISR) (Contract # 200-2010-33976) covering the survey data collection for 2011-2019.

An adequate acknowledgment of all those who made contributions at NCHS and at ISR to the design, conduct, and production of the 2017-2019 NSFG public-use files would require several pages. This brief acknowledgment will only name those who made major contributions.

The NCHS NSFG team is currently comprised of Anjani Chandra (NSFG Team Lead and Principal Investigator), Joyce Abma (Contracting Officer Representative for contract with ISR), Gladys Martinez, Kimberly Daniels, Colleen Nugent, and Jennifer Truong Sayers. Casey Copen

was part of the NCHS NSFG team until 2018, Isaedmarie Febo-Vazquez until 2019, and Chinagozi Ugwu until 2020. Further consultation on other statistical matters at NCHS was provided by Van Parsons and Hee-Choon Shin of the Division of Research and Methodology. During the survey period of the 2017-2019 NSFG, NCHS was under the direction of Charles J. Rothwell, followed by Jennifer Madans. The NSFG team at NCHS is housed within the Reproductive Statistics Branch (RSB) in the Division of Vital Statistics (DVS) at NCHS. Delton Atkinson was Division Director for DVS until September 2018, followed by Steven Schwartz. Paul Sutton is Deputy Director of DVS and acting RSB Chief until July 2019 when Isabelle Horon began serving as RSB Chief. Hanyu Ni was the Associate Director for Science for DVS until December 2019.

Key NSFG contract personnel at University of Michigan's ISR, with their NSFG project roles in parentheses, include Mick Couper (Project Director), William G. Axinn (Deputy Project Director), Heidi Guyer (Field Director/Operations Manager), James Wagner (Senior Mathematical Statistician), Peter Granda followed by Trent Alexander (Director of Data Processing), and William Connett followed by Marcus Blough (Director of Information Security). Other significant contributors at ISR include Michael Shove, Stephanie Windisch, Karl Dinkelmann, Maureen O'Brien, Brady West, Patricia Berglund, Heather Schroeder, and Nancy Oeffner.

NCHS is grateful for the support and contributions of all those noted above, as well as the nearly 123,000 NSFG survey respondents since 1973 who have given their time and energy to provide quality information for the survey.

Suggested Citation for NSFG Public-Use Data and Documentation

As part of the federal statistical system, NCHS supports the dissemination of the NSFG public-use data and documentation files at no charge to the public, and users in all research settings are encouraged to work with these data. NCHS appreciates citation when submitting research grant proposals or publishing their analyses because it will alert their audiences to the source of the data.

> National Center for Health Statistics (NCHS). (2020). 2017-2019 National Survey of Family Growth Public-Use Data and Documentation. Hyattsville, MD: CDC National Center for Health Statistics. Retrieved from **http://www.cdc.gov/nchs/nsfg/nsfg_2017_2019_puf.htm.**