Public Use Data File Documentation

National Survey of Family Growth
Cycle 6: 2002

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 2004

TABLE OF CONTENTS

Appendixes

Appendix 1: File Indexes for NSFG Cycle 6 (App1_FileIndexes.pdf on CD-ROM)

Female Respondent File Index

Female Pregnancy (Interval) File Index

Male File Index

Appendix 2: Recode Specifications for NSFG Cycle 6 (App2_RecodeSpecs.pdf on CD-ROM)

Table of contents for Appendix 2

Female Respondent file recode specifications

Female Pregnancy (Interval) file recode specifications

Male file recode specifications

Main text of User's Guide resides in file called UserGuide_2002NSFG.pdf in the "NSFG Cycle 6 User's Guide" folder on the Public Use CD-ROM. The 2 appendixes are included in that same folder as 2 separate PDF files, as named above.

*For **codebook documentation** of all three data files, please see separate PDF files in the "NSFG Cycle 6 Codebook Documentation" folder on the Public Use CD-ROM, or consult the web-based documentation ("Web-Doc") accessible through links on the NSFG website: www.cdc.gov/nchs/nsfg.htm.*

ACKNOWLEDGEMENTS

An adequate acknowledgement of all those who made significant contributions to the design, conduct, and production of this data file would require many pages. This brief acknowledgement will only name some of those who made some of the more important contributions.

The interviewing, data processing, and data file production for Cycle 6 of the National Survey of Family Growth (NSFG), 2002, were conducted by the University of Michigan's Institute for Social Research (ISR), under a contract with the National Center for Health Statistics (NCHS).

At ISR, **Robert M. Groves** was the Project Director, **William G. Axinn** the Deputy Project Director, and **Krishna L. Winfrey** the Director of Operations. **James M. Lepkowski** was the Chief Mathematical Statistician, **Mick Couper** the Director of Questionnaire Development and Programming, and **Erik W. Austin** the Director of Data Processing. **Ann Biddlecom, Karl A. Dinkelmann, Paul C. Schulz**, and **Lynette F. Hoelter** played key roles in the development, testing, and evaluation of the survey instruments as did **Patricia Maher** and **Grant D. Benson** in the interviewer training and data collection phases of the project. The ISR data file and documentation production team was led by **Peter Granda**, and included programmers **Michael Shove, Jenefer M. Willem**, and **I-Lin Kuo**. The imputation programmer was **John Van Hoewyk**.

At NCHS, the NSFG team, responsible for all aspects of NSFG survey design through data file and documentation production, was comprised of **William D. Mosher** (NSFG Project Officer), **Anjani Chandra**, **Joyce C. Abma, Gladys M. Martinez, and Stephanie J. Willson. Kate Brett,** of the NCHS Office of Analysis, Epidemiology, and Health Promotion, also made significant contributions to the production of the file documentation. Consultation on survey design, variance estimation and other statistical matters was provided by **Karen E. Davis** of the NCHS Office of Research and Methodology.

The 2002 NSFG was jointly planned and funded by the following agencies of the U.S. Department of Health and Human Services:

- the CDC's National Center for Health Statistics,
- the National Institute for Child Health and Human Development (NICHD),
- the Office of Population Affairs,
- the Office of the Assistant Secretary for Planning and Evaluation (OASPE),
- CDC's HIV Prevention Program,
- CDC's Division of Reproductive Health,
- CDC's Office of Women's Health,
- the Children's Bureau of the Administration for Children and Families (ACF);
- ACF's Office of Planning, Research, and Evaluation, and
- ACF's Office of Child Support Enforcement.

NOTE:
Data files as complex as these cannot be guaranteed to be free of errors. If you believe you have found an error, or need assistance, please call the NSFG staff at NCHS at (301) 458-4222.

NATIONAL SURVEY OF FAMILY GROWTH, CYCLE 6
PUBLIC USE DATA

Introduction

Interviewing for the National Survey of Family Growth (NSFG), Cycle 6, was conducted from January 2002 to March 2003 by the Institute for Social Research (ISR) under contract with the National Center for Health Statistics (NCHS). In-person interviews were conducted with 7,643 women 15-44 years of age and 4,928 men 15-44 years of age for a total sample size of 12,571.

Additional details on how the survey was designed and conducted may be found on the NSFG web site (www.cdc.gov/nchs/nsfg.htm), as well as in two reports forthcoming in 2005.

> RM Groves, G Benson, WD Mosher, J Rosenbaum, P Granda, W Axinn, JM Lepkowski, A Chandra. Plan and operation of the 2002 National Survey of Family Growth. *Vital and Health Statistics*, Series 1. Hyattsville, MD: National Center for Health Statistics. Forthcoming in 2005.
>
> JM Lepkowski, WD Mosher, K Davis, R Groves, S Heeringa, J VanHoewyk, T Adams, J Willem. Sample Design, Sampling Weights, Imputation, and Variance Estimation in the 2002 National Survey of Family Growth. *Vital and Health Statistics*, Series 2. Hyattsville, MD: National Center for Health Statistics. Forthcoming in 2005.

The data from the NSFG are used by NCHS and other agencies as the basis for reports and studies on fertility, marriage and cohabitation, contraception, and related issues. A list of over 350 published reports and articles using NSFG data from Cycles 1-5 may be found on the survey's web site, at: www.cdc.gov/nchs/nsfg.htm.

A current list of Cycle 6 publications will be maintained at the NSFG web site and updated periodically.

To make the data more widely available, a standardized Public Use CD-ROM has been prepared for distribution. (See sections on "Data Preparation" and "Restricted-Use Files for the NSFG Cycle 6" for further details.) This User's Guide contains a detailed description of the files on the public use CD-ROM and information for using the data. The files consist of:

(a) the original data for the female respondent, female pregnancy, and male respondent files, and
(b) a set of "recodes," or variables that were created from the original data.

The recodes were created to simplify analyses, and are provided for some key variables in virtually every topic. (See File Indexes in Appendix 1 for a list of recodes provided, and see Appendix 2 for specifications for the recodes.)

The **female respondent** file includes: demographic information, pregnancy history and adoption-related information, and marital and cohabitation history. Data on fecundity, birth expectations, contraceptive use, pregnancy wantedness, use of family planning services,

infertility, and other topics complete this very rich data file.

The **male respondent** file includes: demographic information and information on wives, cohabiting partners, recent sexual partners, and contraceptive use, as well as data on infertility, biological and adopted children, birth expectations, and activities with his children, among other topics.

The **female pregnancy (interval)** file contains detailed pregnancy histories and wantedness of pregnancies, as well as selected respondent characteristics.

Web-Based Documentation (WEBDOC)

To make this very complex data file easier to understand, navigate, and use, the documentation for the survey is available to researchers as a Web-based tool to permit easy access to all variables, quick navigation between different sections of the instrument, and searching for key concepts and questions. At the time of public release of these data, the online documentation is available through links on the NSFG Web site.

This interactive version of the documentation allows users to view the overall structure of the public-use data files including all major sections and the variables which they contain. All information included in the codebook that has been provided with previous Cycles of the NSFG is provided in webdoc. It also allows the NSFG staff to post updates, additions, corrections, and other changes to the documentation as they occur. Users can also consult the questionnaires and other supplementary documentation including several formats of the data collection instruments (see section on “Questionnaires”), which illustrate how this survey was conducted in a computer-assisted interviewing (CAI) environment.

The codebook documentation, based on "Webdoc," has also been included on the Public-Use CD-ROM for the user's convenience. Every subsection of each documentation file is provided in pdf format.

Organization and Use of the Data Files

Organization of the data files

The Cycle 6 NSFG data are provided in three files:

	File Name on CD-ROM
Female Respondent file	FemResp.dat
Female Pregnancy (Interval) file	FemPreg.dat
Male file	Male.dat

The **Female Respondent file** contains one record for each of the 7,643 women in the survey and includes most of the information from their interviews. The **female Pregnancy (Interval) file** contains one record for each of 13,593 pregnancies (both completed pregnancies and current pregnancies), and contains information about the characteristics of each pregnancy and method use and wantedness before each pregnancy. That is, in the Female Respondent file the unit of analysis is the woman, and in the Pregnancy file the unit of analysis is the pregnancy or pregnancy interval. The third data file, the **Male Respondent** file, contains one record for each of the 4,928 men interviewed, and includes most of the information from their interviews. In this file the male respondent is the unit of analysis.

The following is a listing of the column locations of the major sections and key variables for all NSFG Cycle 6 Public Use files.

Male Respondent File: Information for each man

Beginning Column #	Items
1	RESPONDENT ID (CASEID)
13	QUESTIONNAIRE DATA (and computed variables): Sections A-K
2622	RECODES and IMPUTATION FLAGS: Sections A-K
2891	WEIGHTS and related variables
2948	DATE OF INTERVIEW and related variables

Female Respondent File: Information for each woman

Beginning Column #	Items
1	RESPONDENT ID (CASEID)
13	QUESTIONNAIRE DATA (and computed variables): Sections A-J
3749	RECODES and IMPUTATION FLAGS: Sections A-J (including SELECTED

	PREGNANCY (INTERVAL) -BASED questionnaire data and recodes)
4837	WEIGHTS and related variables
4894	DATE OF INTERVIEW and related variables

Female Pregnancy (Interval) File: Information for each pregnancy

Beginning Column #	Items
1	RESPONDENT ID (CASEID)
13	PREGNANCY ORDER (NUMBER)
15	QUESTIONNAIRE DATA (and computed variables): from Sections B and E
275	RECODES and IMPUTATION FLAGS: Sections B & E (including SELECTED RESPONDENT-BASED questionnaire data and recodes)
387	WEIGHTS and related variables
444	DATE OF INTERVIEW

Generally, all pregnancy-, delivery-, and birth-specific variables from sections B and E can be found on the pregnancy (interval) file. All respondent-specific variables can be found on the respondent file.

To facilitate analysis based on women, selected pregnancy-specific variables were placed on the Female Respondent file for each of up to 19 pregnancies. (No respondent in Cycle 6 reported more than 19 pregnancies, though space was allowed for 20.) These include recodes for:

-- pregnancy outcome
-- date pregnancy ended
-- year pregnancy ended
-- age of woman at time of pregnancy outcome
-- formal marital status at pregnancy outcome
-- date of conception
-- age of women at time of conception
-- formal marital status at conception
-- wantedness of pregnancy by R at time of conception (Cycle 4 definition)
-- wantedness of pregnancy by H/P at time of conception (Cycle 4 definition)
-- wantedness of pregnancy by R at time of conception (Cycle 5 definition)
-- wantedness of pregnancy by H/P at time of conception (Cycle 5 definition)

In addition, to facilitate analyses based on pregnancies, some key respondent-specific characteristics were included on the pregnancy (interval) file. These include:

Questionnaire data --
-- nativity status (whether born outside U.S.) and date when she came to the U.S. to stay

Recodes --

-- age at interview
-- race and hispanic origin
-- religious affiliation at interview
-- education at interview
-- insurance coverage status at interview
-- poverty level of household's income at interview
-- receipt of public assistance in the last year
-- labor force status at interview
-- metropolitan residence at interview

Combining Data from Female Respondent and Pregnancy (Interval) files

Analyses using the pregnancy (interval) file may require additional information about women from the respondent file, and analyses using the respondent file may require additional information about pregnancies from the pregnancy (interval) file. Using the common identification number (CASEID), and the pregnancy number (PREGORDR), the pregnancy (interval) and respondent files can be merged to produce a file containing both respondent information and pregnancy information. The resulting file can be either respondent-based (up to 7,643 records) or pregnancy (interval)-based (up to 13,593 records). See Section on "Combining Data from Female Respondent and Pregnancy Files Using SAS" for examples of SAS code that will allow you to merge the respondent and pregnancy files either way.

Weights

Analysts should use the sample weights provided. These will permit replication of the nationally representative estimates that appear in published NCHS reports. The final post-stratified and fully adjusted weight (FINALWGT) is located in columns 4873-4890 in the Female Respondent file, columns 423-440 in the Female Pregnancy file, and columns 2927-2944 in the Male file. To yield number in thousands, as often appears in NCHS reports, each sample weight must be divided by 1,000. For example:

```
WGT1000=FINALWGT_WT/1000
```

For further information, see section on "Sample design, estimation procedures, and variance estimation," or consult the Series 2 report cited in the Introduction.

User Services

Questions and comments concerning this data file may be addressed to:

National Survey of Family Growth staff
Reproductive Statistics Branch
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782

or call the NSFG staff at: (301) 458-4222

or email the NSFG staff at NSFG@cdc.gov

The NSFG staff will assist users as much as possible within the constraints of time and staff availability.

Publications from the National Survey of Family Growth, Cycle 6

The National Center for Health Statistics (NCHS) plans to publish a series of reports from Cycle 6. These will be posted as PDF files at the NSFG web site, which is:

www.cdc.gov/nchs/nsfg.htm.

Users should note that PDF files of virtually all reports published by NCHS can be viewed and downloaded from the NCHS web site. Thus, reports from Cycles 1-5 of the NSFG that were published by NCHS are available from the NSFG web site.

In addition, the web site contains lists of reports and articles in scientific journals published by NSFG staff and others.

This practice will continue in Cycle 6: NCHS reports using the NSFG will be posted on the web site; and published journal articles will be listed on the web site. Users of the NSFG data files should check the NSFG web site periodically.

Individuals and educational institutions may obtain single copies of publications free of charge by writing:

Information Dissemination Staff
National Center for Health Statistics
Metro IV Building, Room 5407
3311 Toledo Road
Hyattsville, MD 20782

or calling: 301-458-4222
or emailing: nchsquery@cdc.gov

Questionnaires

There are numerous benefits of computer-assisted interviewing (CAI) for data quality and ease of interviewing, but one of the challenges CAI poses is how best to represent the computer-programmed interview on paper.

The questionnaires for the NSFG Cycle 6 are available in 3 formats:

-- CAPI-Lite format
-- CAPI Reference Questionnaire (CRQ) format
-- Data Collection Instrument format

The first 2 formats are available as PDF files on the NSFG webpage (or upon request from NSFG staff). The last format is available through the NSFG Cycle 6's Web-based documentation ("Webdoc" – see "Introduction" section for further details).

All 3 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 and perspective 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, users can still get a clear picture of how the questions were asked, in what order, and of which respondents. The full male and female interviews (male Sections A-K and female Sections A-J) are contained in 2 PDF files on the Public Use CD-ROM:

C6female_capiliteMar03final.pdf
C6male_capiliteMar03final.pdf

These files are also accessible through links on the NSFG webpage (www.cdc.gov/nchs/nsfg.htm).

CAPI Reference Questionnaire (CRQ) format

The CRQ represents the fully detailed specifications for the interview that NSFG staff provided to the computer programmers who created the instrument using the Blaise software system.

- 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 a particular respondent appear onscreen. In addition, in some instances flow checks are devoted to the creation of a new variable 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. 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 in the field, rather than requiring resolution after data collection. These consistency checks are generally located in the CRQ after the questions they are intended to reconcile, and enabled the interviewer to return to specific questionnaire items and correct them if necessary.
- Use of additional survey aids, such as Show Cards, Help Screens, and the Life History Calendar (female interview only), is noted on individual questionnaire items. For example, if a question-specific help screen (accessible via the F1 key) 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.

The CRQ for each section of the male and female interviews is provided as a separate PDF file on the Public Use CD-ROM and on the NSFG webpage. (For example, AfemC6CRQ.pdf is the CRQ for female Section A.)

In a few instances, the specifications provided in the CRQ do not match precisely the way that the instrument was programmed. The CRQ was generally not modified to reflect the instrument in these instances, but if there were implications for data quality or interpretation, the user will find additional information in the section on "Data Quality."

Data Collection Instruments

In addition to the PDF formats of the survey questionnaires, users also have the opportunity to view an interactive version of the implementation of the female and male instruments through "Webdoc", a documentation tool available through links on the NSFG web site (www.cdc.gov/nchs/nsfg.htm) (see "Introduction" for further information on Webdoc).

This utility uses an eXtended Markup Language (XML) document which converts information programmed into the computer-assisted interviewing instrument into a set of display pages for all sections of the female and male interviews. The pages for each section contain "go to" or skip instructions, question text fills, valid conditions (universe statements), consistency checkpoints (presented as "edit checks" in the CRQ), interviewer instructions, and all response categories. The "go to" instructions are linked so that users can follow the routing of the instrument based on each response category. The Spanish language version of the instrument is also available for each question and a brief User Guide provides a list of operators and symbols used in the utility.

While not a full representation of how the survey instruments would operate in the field, this tool permits users to study the questionnaires from the interviewer's point of view.

Data Preparation

Preparation of the Data Files for Public Use

Persons who participated in the 2002 National Survey of Family Growth (NSFG) were promised that their answers would be kept confidential. To keep this pledge to respondents, the data files have been modified in preparation for Public Use. All directly identifying information has been eliminated from the public use files. In addition, the state and census region of residence have been withheld. All variables on the files that could otherwise be used to identify individuals have been recoded or categorized, 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 addition, as a final step to prevent identification of individual respondents, the values of some variables have been altered for some respondents. That is, some values in the data set are not the actual values reported by the respondents. However, these alterations were carefully designed to give analysts of the data set similar statistical information as those provided by the unaltered responses. In other words, national estimates and causal models are unlikely to be affected by any of the alterations, except for a very small increase in the variance of some statistics.

For information on the issues and techniques related to disclosure limitation and confidentiality, please consult literature such as the following:

Doyle P, Lane J, Theeuwes JJM, and Zayatz LV, editors. 2001. Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies. New York: Elsevier.

Duncan GT, Stokes SL. 2004. Disclosure Risk vs. Data Utility: The R-U Confidentiality Map as Applied to Topcoding. Chance 17(3):16-20.

Feinberg, SE. 2001. Statistical Perspectives on Confidentiality and Data Access in Public Health. Statistics in Medicine 20(9-10):1347-56.

Feinberg SE, MCIntyre J. 2004. Data Swapping: Variations on a Theme by Dalenius and Reiss. Privacy in Statistical Databases, Proceedings: Annals of the New York Academy of Sciences 3050:14-29.

Muralidhar K, and Sarathy R. 2003. A Theoretical Basis for Perturbation Methods. Statistics and Computing 13(4):329-35.

Reiter JP. 2004. New Approaches to Data Dissemination: A Glimpse into the Future (?) Chance 17(3):11-15.

Trottini M, Feinberg SE. 2002. Modelling User Uncertainty for Disclosure Risk and Data Utility. International Journal of Uncertainty, Fuzziness, and Knowledge-Based Systems 10(5):511-27. October, 2002.

Logical inconsistencies and out-of-range values

During fieldwork, logical consistency of data was maintained through "edit checks" that were included in 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 correct the entry. In Cycle 6, the large numbers of edit checks built into the instruments were intended to minimize the need for post-fieldwork data cleaning and editing. In past cycles, where the interviews were done without computers or where computer capacity was more limited, checking consistency and validity of the data relied more on the application of computer programs ("machine editing") after data collection was completed. In addition, out-of-range values are essentially eliminated in Cycle 6 (as in Cycle 5) because valid ranges are specified and programmed into the instrument, and values outside that range are rejected by the computer.

Other aspects of the questionnaire designed to maximize consistency during data collection were: 1) "summary screens" before 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 chronology of events.

In Cycle 6, the process of checking for consistency by NCHS and ISR staff was focused primarily on the recoded variables and variables related to 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, given the size and complexity of these data files, they cannot be guaranteed to be free of such occurrences.

Two reasons for inconsistencies in the data should be noted:

1) There was an edit check programmed in the instrument, but it was overridden by the interviewer. The programmed edit check may not have been applicable to the respondent's situation; or the interviewer may have misunderstood the answer or the edit check, and overrode the check in order to proceed with the interview. In either case, the interviewer was trained to enter a comment if appropriate, using the F2 key.

2) There was no edit check programmed. There are several areas in the interview where inconsistency could arise, but it was simply not possible to foresee and specify all the edit checks that might be needed.

For further information on the procedures used for preparing these data files, please consult the Series 1 report cited in the Introduction.

Other-Specify Coding

In Cycle 6 a small number of questions contained items to which respondents could specify a response other than those provided: Section E of the female respondent questionnaire. This section, which is about contraception and wantedness of pregnancies, contained several questions that allowed the respondent to specify a response that the interviewer then typed in:

Questionnaire (CRQ) name
EA-21 SP_OTHRMETH
EB-1 SP_FIRSMETH
ED-7 SP_METHHIST
EG-4 WHATMETH

For SP_OTHRMETH, (EA series), if the reponse was not among the methods already asked about, it can be found in a new variable created for this purpose: "NEWMETH". Responses that indicated a method that was already asked about, were coded as "yes" to that method.

For the remaining "Other-specify" variables, the response was coded as one of the existing categories in the variable to which it corresponds. In no case were there sufficient numbers of "other" methods such that a new category could be created. Thus these responses either were 1) re-assigned to an existing method, or 2) assigned to the "other method" category.

Missing Data

Missing data refers to responses of "don't know" or "refused" that were keyed by the interviewer, indicating that the respondent could not or did not provide an answer to a question. In some instances, a code for "not ascertained" was assigned to a variable to account for incorrect routing through the instrument that was determined after fieldwork was completed. 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

(User note: The "proc format" value labels statements provided as *.sas files on the Public Use CD-ROM for the variables on the NSFG data files do not include labels for the codes for "don't know," "refused," and "not ascertained." This is because the codebook documentation only shows those codes if the variable has cases with those particular values. Please use the above guidelines to assign the proper labels to those values in your analyses.)

Because they are imputed, the recoded variables have no missing data, but the 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 the corresponding recode was imputed.

For example, the female respondent file recode CONSTAT1 is associated with the imputation flag called CONSTAT1_I. CONSTAT1_I has non-zero values for 55 cases, which indicates that 55 cases, out of 7,643 female respondents, required some form of imputation on CONSTAT1.

Universe Statements ("Applicable" Specifications)

Not all questions were asked of all respondents because not all questions were relevant to all respondents. If a question was not applicable to a particular respondent, the computer program that ran the questionnaire skipped to the next applicable question. Inapplicable questions are coded as blanks. Some computer programs such as SAS read a blank as a non-numeric character or "system missing" value, but others read it as a zero. Analysts should take care to distinguish between non-numeric values and zeroes in programs used with these data.

The meaning of statistics based on responses to a question will depend, of course, on who was asked the question and who was not. Because of the rather complex skip patterns used in the interview, it is often not apparent who was asked a particular question. In the codebook documentation, the "applicable specifications" or "universe statement" for a question indicates which respondents were asked the question, based on the instrument routing which was driven by the answers to previous questions. The earlier question, and its answer(s) that lead to the question at hand, are included in this universe statement.

For variables whose universes are defined by more than one routing statement, only the routing statement most closely preceding the variable will be described in the universe statement. In that case, the universe statement guides the user to the variable that contains the earlier routing criteria, and this can continue until the full universe statement is accumulated (until the user reaches a variable whose universe statement is "applicable for all respondents").

These other variables referenced in the universe statements are "hyper-linked" in the web-based documentation so that users can go directly to their codebook pages (see Introduction for further information on "Webdoc"). In the codebook files provided in PDF format on the Public Use CD-ROM, users should still find it straightforward to find the relevant codebook entries for variables referenced in the universe statement. To make this easier, question numbers precede the names of all raw/asked variables, names of Blaise-computed variables appear in lower case, and names of recodes appear in upper case.

In addition to consulting the universe statement or "applicable specification" in the codebook documentation, users may also wish to review 1) the routing statements, or "Flow Checks", and 2) the sequence of questionnaire items before and after an item of interest in the CAPI Reference Questionnaire (CRQ, see section on Questionnaires).

The codebook documentation also provides universe statements for the recodes and computed variables (variables constructed from items in the questionnaire). For further information on the recodes, the analyst may wish to examine the Recode Specifications in Appendix 2 of the User's Guide. The definitions of computed variables, beyond what appears in the codebook, are provided in the CRQ.

Recodes

(also see "Imputation" section and Appendixes 1 and 2 – File Indexes and Recode Specifications)

NCHS produces a number of "recoded variables," or "recodes," which are frequently used in NCHS reports. These variables promote the comparable measurement of complex concepts, and make the data file easier for non-NCHS analysts to use

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 fairly simple, while others are quite 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, or Current Contraceptive Status.)

Before using the original data items or constructing their own summary variables, analysts are encouraged to check the File Indexes in Appendix 1 or the Recode Specifications in Appendix 2 to see if a relevant recode exists. All recodes have been edited thoroughly (checked against related data items for consistency); cases that have missing data on a recode have been imputed using a sophisticated multiple regression procedure, the imputed values have been checked for consistency, and imputation flags are provided to indicate whether imputation occurred, and if so, which of two basic types of imputation were used. Published NCHS reports use these recodes whenever available, as they permit internally consistent estimates.

The recodes are clustered together in the variable indexes for each of the 3 data files. Recodes can be distinguished in the codebook documentation by the word "recode" appearing at the end of the question text. Appendix 2 of the User's Guide contains the full specifications for all recodes, roughly in order of the questionnaire sections on which they are based. Consult these specifications for definitions of all code categories, and for a description of the universe for which the recode was applicable. The description of the universe is written to describe cases that were <u>in</u>applicable, for consistency with programming of recodes from prior cycles. If there was a Cycle 5 (1995) equivalent of a Cycle 6 female recode, or if there is a female equivalent of a male recode, it is indicated in the specifications as well.

Imputation

With rare exceptions that are documented where they occur, all **recodes** were imputed for cases with missing data. **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 record which kind of imputation was used.

The most frequently used imputation flag has the following values:

0=Questionnaire data (no imputation)
1=Multiple regression imputation (used most often)
2=Logical imputation

All values other than 0 indicate that the case was imputed for this recode. The definition of each recode in Appendix 2 ("Recode Specifications") includes mention of other recodes that were used to compute it.

Imputation flags are listed in the file indices for each of the three data files. In the Male and Female Respondent File, each questionnaire section's imputation flags follow that section's recodes (i.e., Section A recodes, Section A flags, Section B recodes, Section B flags, etc.). In the Female Pregnancy File, all flags follow all recodes.

The main purpose of imputation was to allow NCHS to produce internally consistent national estimates in its pre-planned Advance Data and Series 23 reports. Actual reported information was never replaced by an imputed value unless the information was obviously incorrect based on other questionnaire data. **We recommend that analysts use the imputed cases for most analyses**. Using weighted data and imputed cases will enable the analyst to replicate results that appear in NCHS reports. However, it may also be desirable for some analyses to be able to examine the impact of imputation, and the flags allow analysts to do that.

The following summary briefly describes the imputation procedure used in the NSFG, Cycle 6. The summary was adapted from the Series 2 report cited below. More complete information on the imputation procedure used in Cycle 6 can be found in this Series 2 report, which will be accessible soon on the NSFG webpage (www.cdc.gov/nchs/nsfg.htm):

> James Lepkowski, et al. National Survey of Family Growth, Cycle 6: Sample design, sampling weights, imputation, and variance estimation. Vital and Health Statistics, Series 2, forthcoming in 2005. Hyattsville, Maryland: National Center for Health Statistics.

The frequency of missing values for most of the recoded variables in Cycle 6 was low, in part because of the use of CAPI, which requires a response before proceeding to the next question. The CAPI program automatically routes the interviewer to the next appropriate question. The program also performs range and consistency checks to rule out logically improbable answers. Fewer than 1 percent of all cases had missing data on most recodes.

In all previous NSFG cycles, the recodes for family income (TOTINCR) and poverty level income (POVERTY – percent of poverty level income based on family size and total

family income) had the largest amounts of missing data---typically about 10-15 percent of all respondents. For example, in Cycle 5, 11.4 percent of respondents had missing data on the POVERTY recode. In Cycle 6, 7.6 % of females and 8.4% of males –about 8 percent overall-- had missing data on the POVERTY recode.

This lower level of missing data in Cycle 6 than in Cycle 5 may be related to the fact that the income questions were put into the self-administered (Audio CASI) part of the questionnaires in Cycle 6, so that the respondent did not have to report them to an interviewer. Another factor that may have lowered the level of missing data on income is the use of two "DK follow-up" questions to narrow the range of family income when respondents said they did not know their total family income. Respondents were asked if their family income was "$20,000 or more," and if they answered "yes" to that question, they were asked if their family income was "$50,000 or more."

As with other recodes in Cycle 6, the income recodes TOTINCR and POVERTY were imputed by a multiple regression technique. In addition, the responses to the two DK follow-up questions were used to guide imputation of these recodes.

For more information on variables (raw and recode) with relatively large or unexpected amounts of missing data, see the section on "Data Quality."

Two main methods of imputation were used for NSFG Cycle 6. The methods differed based on the level of sophistication of the imputation procedure and the availability of data for the imputation. An overview of these methods is provided below. A review by NSFG staff experts was performed after each imputation procedure to evaluate the imputations. If necessary, other values were imputed if the initial imputed values were inconsistent or out of range.

Method 1: Logical imputation -

For some complex recodes with small numbers of missing cases, logical imputation was used. In this procedure, a subject-matter expert at NCHS looked at the values of variables that were related to the variable that was missing, and assigned a logically reasonable value for the missing variable. That is, in the NSFG interview, the response to questions related to the missing value provided sufficient information to assign a consistent "educated guess" for a recode with missing data. Logical imputation was generally limited to variables with very few missing values (e.g., less than 20).

Method 2: Multiple regression imputation - most frequently used method of imputation

The method of imputation used in Cycle 6 uses a multiple regression equation to predict (and impute) missing values for continuous variables; it uses a logistic regression equation to predict (and impute) missing values for discrete or categorical variables. In very simple terms, the dependent variable in the equation is the variable with missing data. The independent variables are all other variables in the data set. In addition, the NSFG and ISR staff worked together to specify and program constraints on the imputed values, to ensure that the imputed values were consistent with other relevant variables, and were "legal" codes on the variable in

question.

The process worked this way:

- the "constraints" were drafted and reviewed by NCHS and ISR;
- the program was written and tested;
- the imputation was run;
- the imputed values were checked for consistency with related variables.
- Inconsistencies were noted, the constraints were revised and improved, and the imputation was repeated.

The imputation theory and method are described in more detail in the following 3 references:

James Lepkowski, et al. National Survey of Family Growth, Cycle 6: Sample Design, Weighting, and Variance Estimation. Vital and Health Statistics, Series 2. Forthcoming in 2005. Hyattsville, MD: National Center for Health Statistics. (available soon at www.cdc.gov/nchs/nsfg.htm)

John Van Hoewyk. 2003. IVEware: Imputation and Variance Estimates Software. The Survey Statistician, July 2003, pages 4-14.

TE Raghunathan, JM Lepkowski, J Van Hoewyk and P Solenberger. 2001. A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. Survey Methodology 27 (1): 85-95, June, 2001.

Abortion Under-Reporting in the NSFG

Abortions have always been under-reported in the National Survey of Family Growth (NSFG) and virtually all other demographic surveys. This has been determined by comparing NSFG weighted estimates of abortions with external data from abortion providers.

Numerator = Weighted number of abortions reported by NSFG respondents in a recent period such as 1996-2000 (2002 NSFG)

Denominator = Number of abortions reported in those same calendar years, based on data reported to CDC's Division of Reproductive Health, and surveys of abortion providers conducted by the Alan Guttmacher Institute.

Using this simple comparison, the estimated percentage of abortions reported by women 15-44, according to NSFG survey year, is shown below.

Cycle & survey year	Percent reported in the NSFG	
Cycle 2 (1976)	45%	
Cycle 3 (1982)	48%	
Cycle 4 (1988)	35%	
Cycle 5 (1995)	45%	(interview)
Cycle 5 (1995)	59%	(Self-administered questionnaire)
Cycle 6 (2002)	**43%**	**(for abortions in 1996-2000)**

(For further information on NSFG Cycles 2-5, see table 1 of H. Fu et al, "Measuring the extent of Abortion Under-reporting in the 1995 National Survey of Family Growth," Family Planning Perspectives 30(3):128-133 and 138, May/June 1998)

CONCLUSION:
As in previous cycles, the NSFG staff advises NSFG data users that, generally speaking, **NSFG data on abortion should not be used for substantive research**.
The NSFG abortion data can be used for:
(1) methodological studies of factors affecting abortion reporting.
(2) studies of contraceptive efficacy, but only after the data are adjusted for the under-reporting of abortion.
The study of the determinants and consequences of abortion is particularly problematic and is, generally speaking, not advised.

Data Quality and Item-Specific Notes for Analysis

As measured by amounts of missing data and inconsistent data, data quality in the 2002 NSFG is high, as in previous cycles. This high quality was obtained through:

-- thorough questionnaire design work, including specification, pretesting, and cognitive laboratory testing;
-- continued use of computer-assisted personal interviewing (CAPI), which allows consistency checks to be built into the interview and eliminates errors related to post-interview data entry;
-- extensive interviewer training; and
-- the use of $40 token of appreciation for sampled women and men who completed the interview, which has been shown in diverse studies to improve response rates and reduce bias in survey data.

Among the thousands of variables in the file, this section notes variables for which further explanation may be helpful for the user, either to (a) provide additional explanation corresponding to selected questionnaire items or recodes, or (b) highlight cases that have questionable or "not ascertained" responses. The issue of questionable or not ascertained data affects only a small percent of cases. Knowing these issues, however, allows users to exclude those cases, impute them, or deal with them in some other way.

As noted earlier:
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 assistance apart from what is discussed below, please call the NSFG staff at NCHS at (301) 458-4222 or email them at nsfg@cdc.gov.

For further information on abortion under-reporting in the NSFG, please see that separate section of the User's Guide.

FEMALE RESPONDENT FILE:

Incorrect routing into questions on Marital Dissolution (CB series) – There was an error in the interview specifications that resulted in a significant number of respondents (more than 500) being mistakenly skipped past questions on how and when their marriage(s) ended. In particular, Flow Check C-14 in the Section C CAPI Reference Questionnaire (CRQ, see section on "Questionnaires") sent respondents to Flow Check C-22 instead of Flow Check C-16, so that they were not asked questions CB-19 MARENDHX through CB-22 WNSTPHX_M/Y. As a result of this routing error, a larger number of women than expected had to have recode values imputed for MARDIS01-05 (century month when marriage ended), MAREND01-05 (how the marriage ended), and the recodes indicating months elapsed between marriage dissolution and other key dates (MAR1DISS, DD1REMAR). This routing error also resulted in more cases than expected needing imputation on pregnancy file recodes based on marriage dissolution dates (FMAROUT5, FMARCON5, and RMAROUT6).

This problem means that studies of marital dissolution and marital status of pregnancies will have larger numbers of imputed cases than in previous NSFG cycles. The problem will be fixed in the next cycle of the NSFG.

Incomplete assignment of CHPNAME fill (name fill for current husband or cohabiting partner, defined in Section C) – The "chpname fill" was used to "fill" the question wording throughout the female interview with the appropriate name or initials for the respondent's current husband or cohabiting partner. This "fill" should have had a non-blank value for all respondents who were currently married or cohabiting; in those few cases where the respondent did not wish to give a name or initials, "your husband" or "your partner" was used for the "fill" in subsequent questions. Due to an error in some flow checks in the CB and CC series, "chpname" remained blank for about 75 married or cohabiting women (that is, for some women with AB-1 MARSTAT = 1 or 2).

In later questions that were meant to be asked for married or cohabiting women, the routing statements were based on "chpname not equal to blank." Since chpname was erroneously blank for these women, these women were erroneously skipped past some questions they should have been asked. For example, they were not asked all applicable questions in Section G on birth expectations. Recodes based on these questions, such as INTENT, were imputed for these women.

Mistaken definition of "sterclin" in Section D – The Blaise-computed variable sterclin was meant to indicate whether the respondent received a sterilization operation at a clinic site within the past 12 months. However, this variable was misspecified in Flow Check D-13. This problem in the definition of sterclin contributed to the Section F problems described below. The sterclin variable is not included on the Public Use File, and the corrected version called R_stclin

is provided as an "intermediate" variable located near the Section F recodes.

DD series on vasectomies of former husbands and cohabiting partners – This series was intended to capture basic information on vasectomies and vasectomy reversals for the respondent's former husbands and cohabiting partners, but the series did not work as intended. Women could report vasectomies for any former husband or former cohabiting partner, but the interview program did not successfully match this vasectomy information with the former husband or partner to whom it applied. Because this information proved unusable as output to the data files, most of the variables from the DD series are not included on the Public Use File. The user will only find the original "filter" question DD-1 VASANY that indicates whether any of the respondent's former husbands or cohabiting partners ever had a vasectomy. The variables in this series that could not be included on the Public Use File would have given the user greater analytic potential for the prevalence and correlates of vasectomy as a contraceptive choice, and it is expected that this series of questions will be corrected for the next NSFG cycle.

About 200 cases skipped past EB-6 USEFRSTS erroneously. This resulted in too many respondents inapplicable on usefstsx (computed) and mthfstsx (computed). This had implications for the recode SEX1MTHD1, as it depends upon mthfstsx.

The respondents skipped erroneously had all used their first method before their first sex. Additionally, the month of first method use was the same as the month of first sex. (This is not contradictory – it means that the first method use occurred before first sex, but both occurred in the same month).

The information on first method ever used, EB-1 FIRSMETH01-04, was intact for many of these respondents. Therefore, for respondents whose first method was a "continuous" method, the recode SEX1MTHD1-4 was assigned that method. This was considered a safe assumption since first sex occurred so close in time to this first (continuous) method use. These methods were:

pill
IUD
depo
partner's vasectomy

Error in Flow Check F-13 (Section F) -- Flow Check F-13 was based on computed variable sterclin, which (as noted above) was not defined correctly. This resulted in too few cases being routed into the questions about the clinic provider for these sterilization operations. In addition, this flow check referenced computed variable anyfster (any female sterilization operation ever) instead of fstrop12 (female sterilization operation within last 12 months), which resulted in too many women being asked the subsequent questions, but the *net* effect of these 2 errors in Flow Check F-13 is that *too few* women were asked the appropriate questions about the clinic provider.

Error in Flow Check F-17d (Section F) -- Section F of the female questionnaire was trimmed down between Cycle 5 and Cycle 6 in order to reduce the overall length of the questionnaire. In this process, the section on "ever received family planning services" was deleted, along with some of the details about the first service received. A new series (FB) was created to collect some of the information on the first service received. This new series of questions attempted to route respondents without knowing if they had ever received a service. To compensate for this absence of information, information on "ever use" of family planning methods was used from section E. Flow Check F-17d in this new series routed respondents whose first service was not in the last 12 months to FB-7 BCPLCFST (where R received first service) as was done in the 1995 NSFG. But because information about 1st service was not collected earlier in the 2002 questionnaire, respondents should have been routed to FB-6 FSTSERV (service(s) received the first time) before being asked FB-7 BCPLCFST (source of 1st service). Therefore, the information in FSTSERV1 is not complete because about a third of eligible respondents (about 800 sexually active 15-24 year olds) were skipped past this variable. As a result of this error in Flow Check F-17d, the FSTSERV1 variable has not been included on the Public Use data file.

FC-2 KNDMDHLP -- Question FC-2 KNDMDHLP, "What kind of medical service did you receive at the clinic" was specified to allow multiple reporting of services. There was a problem in the instrument, and the question only allowed the reporting of one service.

GB-3 JINTENDN -- Respondents who are married or cohabiting were only asked GB-3 JINTENDN (number of children they intend) if they answered yes to GB-1 JINTEND (whether they intend to have another). The questions for those who are not currently married should have been worked the same way but this was not done. Respondents who were not currently married and who answered either yes or no to GC-1 INTEND, were also asked GC-3 INTENDN (number of children they intend). Therefore, when looking at number intended researchers need to control for INTEND=yes to make the two variables compatible.

Erroneous date of interview (cmintvw) -- One female respondent has January 2002 as her date of interview (cmintvw = 1225). This is an erroneous value for cmintvw; her interview actually took place in October of 2002. We have not altered this date of interview because many key questions referenced "the year before the interview," and this respondent was asked about the period "since January 2001" (1 year before January 2002). Altering the interview date for this case would have resulted in inconsistency between the cmintvw value that was used to guide routing and question wording in the interview, and the values of both raw data and recoded variables.

FEMALE PREGNANCY (INTERVAL) FILE:

131 women were skipped past EG-10 TIMINGOK erroneously, for their pregnancies --

These cases are coded "7" (not ascertained) on this variable. Women who responded "not sure, don't know" to the question (EG-6 WANTBOLD) "Right before you became pregnant (this time) did you yourself want to have a(nother) baby at any time in the future?" and then responded "probably yes" to the question (EG-7 PROBBABE) "It is sometimes difficult to recall these things but, right before this pregnancy began, would you say you probably wanted a baby at some time in the future or probably not?" were skipped past this question on pregnancy timing.

MALE FILE:

Biological children in chronological order

In the male questionnaire, respondents were asked about their biological children in the context of questions about the mothers of these children. Based on consultation with experts in surveys of men, this approach was considered to yield the most accurate reporting of men's sexual, fertility, and contraception experiences. However, with this approach, questionnaire information on a biological child is located on the data files in the section where a man reported the child:

- In Section C if the child's mother is his current wife or cohabiting partner
- In Section D if the child's mother was his last sexual partner ever, or one of his 3 most recent partners in the last 12 months
- In Section E if the child's mother was a former wife or his 1st premarital cohabiting partner (and not a recent sexual partner)
- In Section F if the child's mother was any other sexual partner not discussed in Sections C-E

To assist users who wish to analyze information on men's biological children based on chronological order of birth, the male data file includes selected variables derived from the section-specific biological child variables in Sections C-F. These chronologically ordered variables are located on the male file at the end of Section F questionnaire items. The variables are arranged as arrays of 10 variables each because no respondent reported more than 10 biological children. The variables on the Public Use File include:

- BIODOB1-10: Century month of child's birth
- BIOSEX1-10: Sex of child
- BIOAGE1-10: Age of child
- BIOHH1-10: Whether child lives in same household with respondent
- BIOMAR1-10: Whether respondent was married to child's mother at time of child's birth
- BIOCOHB-10: Whether respondent was living with child's mother at time of child's birth (includes cohabiting or married)
- BIOLRNPG1-10: When respondent learned of the pregnancy (before or after child was born)
- BIOLGPAT1-10: Whether legal paternity was established (if unmarried at child's birth)
- BIOHSPAT1-10: Whether paternity was established at the hospital
- BIOLVEVR1-10: Whether respondent ever lived with child (if not living with child now)
- BIOHWFAR1-10: How far away child lives (in miles) from respondent
- BIOWANT1-10: Wantedness of the pregnancy by respondent

- BIOHSOON1-10: Timing of the pregnancy
- BIOHPYPG1-10: Respondent's happiness about the pregnancy

The table below illustrates how these chronologically arranged variables are derived from questionnaire items in Sections C-F. Please consult the codebook documentation to see further details on universe statements and response categories for these variables or for the questionnaire variables on which they were based.

	Source variables in ...			
Chronological variable	**Section C (CG series)**	**Section D (DH series)**	**Section E (ED series)**	**Section F (FA series)**
BIODOB[x]	CWPCHDOB_M/Y -- cmchdob(nnn)	PXCXBORN_M/Y -- cmchdob(nnn)	FWPCHDOB_M/Y -- cmchdob(nnn)	OBCDOB_M/Y -- cmchdob(nnn)
BIOSEX[x]	CWPCHSEX	PXCXSEX	FWPCHSEX	OBCSEXX
BIOAGE[x]	based on cmchdob	based on cmchdob	based on cmchdob	based on cmchdob
BIOHH[x]	CWPCHLIV	PXCXLIV	FWPCHLIV	OBCLIV
BIOMAR[x]	CWPCHMAR	PXCXMARB	FWPCHMARB	not applicable
BIOCOHB[x]	CWPCHMAR & CWPCHRES	PXCXMARB & PXCXRES	FWPCHMARB & FWPCHRES	OBCMLIV
BIOLRNPG[x]	CWPCHLRN	PXCXKNOW	FWPCHLRN	OBCKNOWX
BIOLGPAT[x]	CWPCHLEG	PXCXLAW	FWPCHLEG	OBCLAW
BIOHSPAT[x]	CWPCHHOP	PXCXHOP	FWPCHHOP	OBCHOP
BIOLVEVR[x]	CWPCHEVR	PXCXEVER	FWPCHEVR	OBCEVER
BIOHWFAR[x]	CWPCHFAR	PXCXFAR	FWPCHFAR	OBCFAR
BIOWANT[x]	CWPCHWNT	PXWANT	FWPRWANT	OBCRWANX
BIOHSOON[x]	CWPCHSON	PXSOON	FWPSOON	OBCSOONX
BIOHPYPG[x]	CWPCHHPY	PXHPYG	FWPHYPG	OBCHPYX

In summary:
These chronologically variables are essentially identical in content to the source variables in Sections C-F. Users whose primary goal is to examine data on men's biological children in order of their birth should use these chronologically arranged variables. If, however, their primary goal is to examine men's fertility in the context of their relationships with their children's mothers, it may be more appropriate to use the source variables in Sections C-F

Map to variable names for contraceptive method use experiment in the Male CRQ & data file

Since the questions on contraceptive method use in Sections C and D of the male questionnaire were involved in an experiment to test the effects of question wording, there are different sets of questions in these 2 sections that are analogous to each other. The following is a list of the questions and their descriptions to help users, should they want to use the raw data items.

Method use with current wife or cohabiting partner (Section C):

30% experimental group:

CRQ	Data file	Description
CE-3 CWPLUSE	CWPLUSE	R or CWP use a method at last sex?
CE-4 CWPLMET	CWPLMET01-04	method R or CWP used at last sex :1st-4th

70% experimental group:

CRQ	Data file	Description
CE-5 CWPLUSE1	CWPLUSE1	R use method at last sex with CWP?
CE-6 CWPLMET1	CWPLMET11-14	Method R used at last sex with CWP:1st-4th
CE-7 CWPLUSE2	CWPLUSE2	CWP use method at last sex with R?
CE-8 CWPLMET2	CWPLMET21-24	Method CWP used at last sex with R:1st-4th

Method use with nonmarital / noncohabiting 3 last (recent) partners (Section D)

30% experimental group:

CRQ	Data file	Description
DD-3 PXLUSE	PXLUSE	R or P use method at last sex? (last partner)
DD-4 PXLMETH	PXLMETH01-04	Meth R or P used last sex:1st-4th method, last P
DD-3 PXLUSE2	PXLUSE2	R or P use method at last sex? (2nd-to-last partner)
DD-4 PXLMETH	PXLMETH11-14	Meth R or P used last sex:1st-4th method, 2nd-to-last P
DD-3 PXLUSE3	PXLUSE3	R or P use method at last sex? (3rd-to-last partner)
DD-4 PXLMETH	PXLMETH21-24	Meth R or P used last sex:1st-4th method, 3rd-to-last P

70% experimental group:

CRQ	Data file	Description
DD-5 PXLRUSE	PXLRUSE	R use method at last sex with P? (last partner)
DD-6 PXLRMETH	PXLRMETH1-4	Method R used last sex with P:1st-4th (last partner)
DD-7 PXLPUSE	PXLPUSE	P use a method at last sex w/R?(last partner)
DD-8 PXLPMETH	PXLPMETH1-4	Method P used last sex with R:1st-4th(last partner)
DD-5 PXLRUSE2	PXLRUSE2	R use method at last sex with P? (2nd-to-last partner)
DD-6 PXLRMETH	PXLRMETH5-8	Method R used last sex with P:1st-4th (2nd-to-last P)
DD-7 PXLPUSE	PXLPUSE2	P use a method at last sex w/R?(2nd-to-last P)
DD-8 PXLPMETH	PXLPMETH8-11	Method P used last sex with R:1st-4th(2nd-to-last P)
DD-5 PXLRUSE	PXLRUSE3	R use method at last sex with P? (3rd-to-last partner)
DD-6 PXLRMETH	PXLRMETH9-12	Method R used last sex with P:1st-4th (3rd-to-last P)
DD-7 PXLPUSE	PXLPUSE3	P use a method at last sex w/R?(3rd-to-last partner)
DD-8 PXLPMETH	PXLPMETH15-18	Method P used last sex with R:1st-4th (3rd-to-last P)

The following recodes on the Public Use File combine these experimental sets of questions, and generate measures of contraceptive method use: LSEXUSE1-LSEXUSE4; METH12M1-METH12M4; METH3M1-METH3M4; SEX1MTHD1-SEX1MTHD4. See Appendix 2 for further information on how these recodes were defined.

Error in routing in Section G -- At the beginning of the fieldwork period in Spring 2002, an error was discovered in the CAPI programming for respondents with non-coresidential children under age 5. These respondents were skipped past the questions on activities with non-coresidential children (GB-12 NCFEED thru GB-15 NCREAD). This problem was fixed in the field but it had already affected 52 cases. The activity questions affected by this problem were coded 7 to indicate they were "not ascertained."

Recode LSEXRAGE – On the male recode LSEXRAGE (respondent's age at last sexual

intercourse), 4 respondents have a value of 45 years. Three of these 4 respondents were 44 years old at the time of the screener (computed variable agescrn=44), and had their 45th birthday before the interview was conducted (computed variable age_r=45), so LSEXRAGE is correctly equal to 45.

For one of the 4 respondents, his age at last intercourse LSEXRAGE was actually 44, but was computed to be 45 based on the fact that his last intercourse occurred in the month of interview, which happened to be his birth month. This respondent reported in question AA-1 AGE_A that he was 44 years old, indicating that the interview occurred before his actual birthday in that month. However, the formula for computing LSEXRAGE (century-month of last sexual intercourse minus century-month of respondent's birth, divided by 12) yielded 45 years.

Inconsistency between "3 most recent sexual partners in the last 12 months" information reported in Section B and Section D:

In Section B of the male questionnaire, respondents were asked to list up to 3 most recent sexual partners in the last 12 months, starting with the most recent, then the 2nd most recent, then the 3rd most recent. Later, in Section D, respondents were asked to report their dates of last sex with each of these partners.

The variables involved in this issue are the following 3 raw variables from Section B, not included on the Public Use File, but used in routing through the instrument:

BD-1 P1NAME (name or initials of the most recent sexual partner)
BD-7 P2NAME (name or initials of the 2nd most recent sexual partner)
BD-13 P3NAME (name or initials of the 3rd most recent sexual partner)

and these 3 computed variables in Section D:

cmlsxp (date of last sex with the most recent sexual partner)
cmlsxp2 (date of last sex with the 2nd most recent sexual partner)
cmlsxp3 (date of last sex with the 3rd most recent sexual partner)

Most respondents reported dates of last sex in Section D that were consistent with the ordering of partners in Section B. That is, the date of last sex with the *most recent* partner (cmlsxp) was indeed later than the dates of last sex with the other partners they reported in Section B (cmlsxp was later than both cmlsxp2 and cmlsxp3). However, in a little over 100 cases, the date in cmlsxp was not the most recent date, compared to cmlsxp2 and cmlsxp3. This means that there was a contradiction between the information given in Section B and in Section D.

The recode LSEXDATE2 assigns date of last sex according to the most recent date. LSEXDATE assigns the date according to the partner he reported was his most recent partner, regardless of the actual date given later. Other recodes that are based on date of last sex/last partner use LSEXDATE and not LSEXDATE2.

A variable was created to help users by giving them an option of selecting the partner who was the most recent, based on either criterion. (The Section B information or the Section D information). This variable is called **orderflag** (an "intermediate" variable on the data file, located with the Male recodes)**,** and it identifies cases with out-of-order Section D partner dates (cmlsxp, cmlsxp2, cmlsxp3). Note that the inapplicable cases on this variable include those who are truly inapplicable (for example, never had sex), as well as those lacking valid dates. This flag is to help handle the valid data, rather than to distinguish the applicable universe from those "erroneously skipped".

Codes and value labels for orderflag:

. = inapplicable or no valid information on partner dates
1= Section D partner dates (nonmissing) are in proper chronological order, or only one valid date was reported
2 = Section D partners reported out of order: affects last partner and possibly others
3 = Section D partners reported out of order: affects 2nd-to-last and 3rd-to-last partners only

Error in routing into DB-3 LIVTOGN – The question DB-3 LIVTOGN was intended to ask about premarital cohabitation for each of the respondents (up to 3) most recent sexual partners in the last year, but an error in flow checks leading up this question resulted in *too few* respondents getting asked this question. Not all respondents who were ever married to their recent partners were asked if they lived together before they got married. As a result of this error, some of the recodes related to cohabitation may have been defined incorrectly for some of these respondents.

For example, the recode COHAB1 is meant to represent the date of men's first cohabitation outside of marriage. If a respondent's earliest cohabitation was actually a premarital cohabitation with one of his recent sexual partners, it is possible that COHAB1 was given a value that was too late. This is somewhat unlikely, however, because the only values COHAB1 could have been given in this case would have come from the start of his premarital cohabitation with a current wife (in Section C). It is unlikely that a currently married respondent who cohabited premaritally with his current wife would have also had a former wife with whom he had sex in the last 12 months *and* with whom he cohabited premaritally.

Nonetheless, LIVTOGN was erroneously skipped for all recent partners to whom respondents were ever married, and the number of cases who are "inapplicable" in the codebook documentation is higher than expected due to this routing error. The recode COHEVER, which indicates whether the respondent *ever cohabited outside marriage,* may *slightly underestimate* the true prevalence of this behavior for men 15-44 in 2002.

Computed variable totpregs_c in Male Section F – The computed variable totpregs_c was intended to indicate the sum of all pregnancies respondents reported fathering, throughout the interview. The specifications for this variable were incomplete, and did not properly account for "don't know," "refuse," and system-missing values on some of the source variables. This resulted in a few (less than 10) extreme or implausible values, which then yielded implausible

values on certain recodes. Because of this problem with totpregs_c, these cases were imputed on the affected recodes, including COMPREG (total number of completed pregnancies ever fathered).

Attitude question JG-3 SAMESEX – There was an error in the wording of the attitude question JG-3 SAMESEX in the male questionnaire. The text of the question should have said "Sexual relations between two adults of the same sex are all right" to match the female version, but instead it read "Sexual relations between two adults of the same sex are always wrong." This makes the male and female versions of this question no longer comparable. This problem will be fixed in Cycle 7.

Restricted-Use Files for NSFG Cycle 6

When the National Center for Health Statistics (NCHS) collected the data from respondents to the National Survey of Family Growth (NSFG), those respondents were promised that the information they provided would be held "strictly confidential." The NCHS is legally required to keep that promise. In order to do so, a number of things were done to the data files to prevent disclosure of the identities of the respondents, and at the same time attempt to preserve all the analytical value of the data. (see section on "Data Preparation" for further information)

The public use file (included with this documentation) was reviewed by the NCHS Disclosure Review Board and the NCHS Confidentiality Officer. In response to that review, the NSFG staff and contractor eliminated some variables from the public use file, and for other variables, combined or recoded categories identifying very small groups.

Some variables, however, pose a substantial risk of disclosure, and cannot be included on Public Use Files. These are made available to the research community by means of restricted-use data files. This section describes these files in general terms. Interested researchers should contact NCHS for additional information.

The three main groups of variables that comprise restricted-use data files for the NSFG Cycle 6 are:

- Omitted Items
- Interviewer Variables
- Contextual Data

Omitted items

As can be seen in the "Outline of Contents of the Data Files," the NSFG Cycle 6 questionnaires contained a number of items designed to provide a comprehensive description of current and past behavior related to the risk of acquiring sexually transmitted diseases (STD) including the Human Immunodeficiency Virus, or HIV, the virus that causes AIDS. These

questions were asked via Audio Computer-Assisted Self-Interviewing, or ACASI, in which the respondent enters the answers directly into the computer, without the help of an interviewer. The object of ACASI was to give respondents a more private opportunity to report sensitive information.

The omitted items file includes most of the items in ACASI and selected items from the interviewer administered portion of the interview (CAPI). As noted above, the items collected in ACASI (female section J and male section K) pertain mostly to pregnancy reporting, drug use, and STD/HIV-risk behaviors, which are too sensitive for inclusion on the public use files.

Interviewer variables

Interviewer observations of the interview process, as well as interviewer characteristics, comprise another restricted-use file. Much of this material was collected for methodological reasons, to understand better what factors lead to variations in data quality.

Contextual data

Contextual data is information on the context, or environment, in which respondents live (for example, the male or female unemployment rate in the area, or the percent of households with incomes below the poverty level). The 2000 Census Summary Files are the source of some contextual data, at the State, County, Census Tract, and Block Group levels. As in Cycle 5, a contextual data file will be available for the 2002 NSFG. Contextual variables will be based, to the extent possible, on where the respondents lived at 2 points in time – April 1, 2000 (the time of the last U.S. Census) and the date of interview for the 2002 NSFG.

At the time of this writing, the list of variables to be included on the contextual data files for Cycle 6 had not been finalized, but will include some variables from the County and City Data Book at the County level, as well as the Summary file data at the County, Census Tract, and Block Group level. Users can get a rough idea of the kinds of information that may be included by consulting the Cycle 5 contextual variable list (see Series 23, Number 23 by Mosher, Deang, and Bramlett, particularly Appendix II).

Access to the Restricted-Use NSFG Files

While these restricted-use data files cannot be made generally available, the National Center for Health Statistics wishes to promote the scientific purposes for which the data were collected. To this end, the Center can make the data available to researchers for specified scientific analyses, under special arrangements that assure confidentiality and protection of the data commensurate with that provided by the Center itself.

Access to the **contextual data files** will be through the NCHS Research Data Center. For the **"Omitted Items**" and **"Interviewer variables**" files, specific access procedures were still

being determined as this public-use file was being released. Researchers who wish to learn more about or apply for access to any of these files -- Omitted Items, Interviewer Variables, or Contextual Data -- should write on their organization's letterhead to:

National Survey of Family Growth staff
Reproductive Statistics Branch
Division of Vital Statistics
National Center for Health Statistics
3311 Toledo Road, Room 7318
Hyattsville, MD 20782

Researchers may also find useful information in

(a) the Series 23, Number 23 report cited above, available on the NSFG web site: www.cdc.gov/nchs/nsfg.htm and on

(b) the NCHS Research Data Center webpage: www.cdc.gov/nchs/r&d/rdc.htm

Date Codes

During the interview, dates of events were recorded as month and year, except for the respondent's date of birth, which was recorded as month, day, and year.

For inclusion in the data file, month and year for most dates reported in the interview, including the respondent's date of birth, were 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 February 1959 is (59 x 12) + 2= 710.

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 March 2002 is (102 x 12) + 3 = 1227.

The century month form is convenient for computing intervals between dates.

With the exception of one recoded date variable (DATEUSE1 on the female respondent file) that has a leading 9 to indicate that the value was estimated, all date variables in the file are 4 columns long. The following codes were used for the 3 types of missing data on date variables:

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

The century month codes from 301 through 1248 are shown in the array below with the years from 1925 through 2003 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 2002 NSFG were conducted between March 2002 (century month 1227) and March 2003 (century month 1239).

DATE CODES

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1925	301	302	303	304	305	306	307	308	309	310	311	312
1926	313	314	315	316	317	318	319	320	321	322	323	324
1927	325	326	327	328	329	330	331	332	333	334	335	336
1928	337	338	339	340	341	342	343	344	345	346	347	348
1929	349	350	351	352	353	354	355	356	357	358	359	360
1930	361	362	363	364	365	366	367	368	369	370	371	372
1931	373	374	375	376	377	378	379	380	381	382	383	384
1932	385	386	387	388	389	390	391	392	393	394	395	396
1933	397	398	399	400	401	402	403	404	405	406	407	408
1934	409	410	411	412	413	414	415	416	417	418	419	420
1935	421	422	423	424	425	426	427	428	429	430	431	432
1936	433	434	435	436	437	438	439	440	441	442	443	444
1937	445	446	447	448	449	450	451	452	453	454	455	456
1938	457	458	459	460	461	462	463	464	465	466	467	468
1939	469	470	471	472	473	474	475	476	477	478	479	480
1940	481	482	483	484	485	486	487	488	489	490	491	492
1941	493	494	495	496	497	498	499	500	501	502	503	504
1942	505	506	507	508	509	510	511	512	513	514	515	516
1943	517	518	519	520	521	522	523	524	525	526	527	528
1944	529	530	531	532	533	534	535	536	537	538	539	540
1945	541	542	543	544	545	546	547	548	549	550	551	552
1946	553	554	555	556	557	558	559	560	561	562	563	564
1947	565	566	567	568	569	570	571	572	573	574	575	576
1948	577	578	579	580	581	582	583	584	585	586	587	588
1949	589	590	591	592	593	594	595	596	597	598	599	600
1950	601	602	603	604	605	606	607	608	609	610	611	612
1951	613	614	615	616	617	618	619	620	621	622	623	624
1952	625	626	627	628	629	630	631	632	633	634	635	636
1953	637	638	639	640	641	642	643	644	645	646	647	648
1954	649	650	651	652	653	654	655	656	657	658	659	660
1955	661	662	663	664	665	666	667	668	669	670	671	672
1956	673	674	675	676	677	678	679	680	681	682	683	684
1957	685	686	687	688	689	690	691	692	693	694	695	696
1958	697	698	699	700	701	702	703	704	705	706	707	708
1959	709	710	711	712	713	714	715	716	717	718	719	720
1960	721	722	723	724	725	726	727	728	729	730	731	732
1961	733	734	735	736	737	738	739	740	741	742	743	744
1962	745	746	747	748	749	750	751	752	753	754	755	756

	JAN	FEB	MAR	APR	MAY	JUN	JUL	AUG	SEP	OCT	NOV	DEC
1963	757	758	759	760	761	762	763	764	765	766	767	768
1964	769	770	771	772	773	774	775	776	777	778	779	780
1965	781	782	783	784	785	786	787	788	789	790	791	792
1966	793	794	795	796	797	798	799	800	801	802	803	804
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

Sample Design, Estimation Procedures, and Variance Estimation

The following summary briefly describes the sample design, weighting, imputation, and variance estimation for the NSFG, Cycle 6. The summary was adapted from a preliminary draft of the Series 2 report cited below. More complete information on the imputation procedure used in Cycle 6 can be found in that report:

> JM Lepkowski, et al. Sample Design, Sampling Weights, Imputation, and Variance Estimation in the 2002 National Survey of Family Growth. *Vital and Health Statistics*, Series 2. Hyattsville, MD: National Center for Health Statistics. Forthcoming in 2005.

The 2002 National Survey of Family Growth (NSFG) obtained detailed information on factors affecting childbearing, marriage, and parenthood from a national probability sample of 12,571 men and women 15 to 44 years of age. The Series 2 report is the technical documentation of the procedures used to select the sample, develop the sampling weights that permit valid population estimates, impute missing data, and estimate sampling errors. For readers seeking a general understanding of the survey procedures, this section provides a summary of the procedures used.

The National Survey of Family Growth is designed and administered by the National Center for Health Statistics (NCHS), an agency of the Department of Health and Human Services. The purpose of the survey is to produce national estimates of:

- factors affecting pregnancy---including sexual activity, contraceptive use, infertility, and sources of family planning services;
- factors affecting marriage, divorce, cohabitation, and adoption;
- what men and women do to raise their children; and
- men's and women's attitudes about sex, childbearing, and marriage.

For Cycle 6, interviewing and data processing were conducted by the University of Michigan's Institute for Social Research, Robert Groves, Project Director, under a contract with NCHS.

A national probability sample of 12,571 men and women 15-44 years of age in the non-institutionalized population of the United States were interviewed between March 2002 and March 2003. The interviews were conducted in person by trained female interviewers using laptop, or notebook, computers---a procedure called computer-assisted personal interviewing (CAPI). The interviews for women averaged 85 minutes; the interviews for men averaged about 60 minutes.

For women, data were collected on each pregnancy (if any); contraceptive use by her and her partner; her ability to bear children; the use of medical services for contraception, infertility, and prenatal care; marriage and cohabitation history; and a variety of demographic and economic characteristics.

For men, data were collected on marriage histories; contraception, children fathered;

parenthood activities and attitudes; and demographic and economic characteristics.

Men and women were also asked questions on behaviors related to the risk of sexually transmitted diseases, including HIV. Those data are available in the omitted items file (see section on "Restricted-Use Files for NSFG Cycle 6.")

This section describes briefly how the sample was designed and selected, how sampling weights were computed and adjusted to compensate for nonresponse, how sampling errors were estimated, and how missing data were imputed for some data items. For a more detailed account of these topics, see the forthcoming Series 2 report.

Sample Design

A total of 12, 571 men and women were interviewed from a national probability sample of households in the United States. Men and women were selected from 121 Primary Sampling Units (PSU's). A PSU is a metropolitan area, a county, or a group of adjacent counties. PSU's were located in nearly every state, and included all of the largest metropolitan areas in the United States.

From each PSU, secondary units, called segments, were selected. Segments are, roughly, neighborhoods, or groups of adjacent blocks. In each selected segment, addresses were listed, and a sample of the addresses was taken. The sampled addresses were contacted, and a "screener" interview was attempted, in which the persons living at that address are listed. If one or more eligible persons (15-44 years of age) were living at that address, one person was randomly selected and asked for an interview.

Sampling Weights

A simple random sample in which response rates and coverage were the same in every sub-group would be a "scale model" of the population. However a survey sample is almost never a scale model in that sense. Groups are often selected at different rates and often have different response rates. For example, in the NSFG, non-Hispanic black men and women account for 19.6 percent of all respondents in the sample but only 12.9 percent of the population 15-44 years of age. "Sampling weights" adjust for these different sampling rates, response rates, and coverage rates so that accurate national estimates can be made from the sample.

A respondent's sampling weight can be interpreted as the number of persons in the population that he or she represents. For example, if a woman's sampling weight is 8,000, then she represents 8,000 women in the population. For the NSFG, the fully adjusted sampling weights were assigned to each respondent and consisted of 4 factors. The first factor is the inverse of the probability that the case was selected. For example, if the probability of section is 1 in 6,000, then the initial sampling weight is 6,000. The second factor is an adjustment for non-response, which was calculated separately based on the probability of completing a screener, and the probability that a completed screener would result in a completed interview. The third factor is an adjustment to control totals of the number of persons by age, sex, race and Hispanic origin, provided by the US Census Bureau. This process is called post-stratification. The fourth factor

is trimming, which reduces the values of a few extremely large weights.

Item Imputation

In any survey, not every question is answered by every person interviewed. Sometimes a respondent cannot remember the fact asked for in a question; sometimes he or she may refuse to answer. Other times, the answer that the respondent gives is clearly inconsistent with other information in the interview, so one or more of the inconsistent answers is set to missing. Such "missing data" create inconsistencies in estimates, which may be confusing for many users of the data. Assigning values to these missing items is called "imputation." Imputation makes the data more complete, more consistent, and easier to use.

In Cycle 6 of the NSFG, there are thousands of variables in the data file. Of these many variables, nearly 400 recoded variables (called "recodes") were selected because they are used frequently in analysis. Missing data for these recodes could create inconsistencies among survey estimates and confusion among data users of both the published data and the micro-data file, so these variables were imputed. Selecting, editing, and imputing these variables was one way to decide which variables should be examined carefully to ensure high-quality data, without unduly delaying the release of the data file.

The frequency of missing values for the recoded variables in Cycle 6 was low, in part because CAPI requires the interviewer to enter an acceptable response and then goes automatically to the next appropriate question. The program also performs range and consistency checks to prevent logically impossible answers. The 2 imputation techniques used in Cycle 6 were:

- Logical imputation
- Regression imputation

Logical imputation involves having a subject-matter expert (usually at NCHS) examine variables related to the variable in question, and assign a value that is consistent with those other variables and is an educated guess of the true value when there is any ambiguity. Regression imputation, as used for NSFG Cycle 6, used software that imputes a missing value using all other variables in the data set as predictors. A major part of the work of imputation involves making certain that the values imputed are within acceptable ranges, and are consistent with other data reported by the respondent.

Except when it was obviously incorrect, actual reported data were never replaced by an imputed value. For each recoded variable in the database, an imputation flag identifies whether the value of that variable was imputed or not. Using the imputation flag, a researcher can identify the observations with an imputed value and the specific type of imputation procedure used for each specific recoded variable.

Variance Estimation

The sampling variance is a measure of the variation of a statistic (such as a proportion or

a mean) caused by having taken a sample instead of interviewing the full population. (For example, in the NSFG, the sampling error measures the variation caused by interviewing 12,571 men and women in the NSFG instead of the 120 million men and women 15-44). It measures the variation of the estimated statistic over repeated samples. The sampling variance is zero when the full population is observed, as in a census.

For the NSFG, the sampling variance estimate is a function of sampling design and the population parameter being estimated, and it is called the design-based sampling variance. The NSFG data file contains a final weight and information necessary to estimate the sampling variance for a statistic. Many statistical software packages, such as SAS and SPSS, will, by default, compute "population" variances, which may severely underestimate or overestimate the sampling variances. Special software is required to accurately estimate sampling errors in a complex sample such as the NSFG, but such software is becoming more and more common, and is easier to use and obtain.

Examples of how to use such software to estimate sampling errors for the NSFG are included in the Series 2 report mentioned above, and are also available through links on the NSFG webpage (www.cdc.gov/nchs/nsfg.htm).

For example, the NSFG Cycle 6 design parameters needed to estimate variances using SAS/SUDAAN software are:

DESIGN = WR (with replacement)

NEST statement:
- SECU = PSU (cluster) variable
  - SECU_R (if using the female respondent file)
  - SECU_P (if using the female pregnancy file)
  - SECU (if using the male file)
- SEST = stratum variable

WEIGHT statement:
- FINALWGT = final post-stratified, fully adjusted weight

Here is one example (based on SUDAAN version 9 code) for a tabulation of recode HADSEX with the male data file, using the DEFF option to calculate design effects:

```
proc sort data=nsfgmale;
by SEST SECU;
proc crosstab data=nsfgmale design=wr deff;
nest SEST SECU;
weight FINALWGT;
subgroup hadsex;
levels 2;
table hadsex;
print nsum wsum rowper serow deffrow;
run;
```

Conclusion

Because of the complex sample design of the NSFG, analysts should use the weights in analysis whenever possible, and use software that will compute "designed-based" estimates of sampling errors. Failure to use the weights and accurate variance estimates may lead to biased or inaccurate findings and conclusions.

File Characteristics

	Number of Records (observations)	Record Length (number of columns)	Number of Variables
Female respondent file File = FemResp.dat	7,643	4,927	3,087
Female pregnancy (interval) file File = FemPreg.dat	13,593	447	243
Male file File = Male.dat	4,928	2,986	1,993

Outline of the Contents of the NSFG Cycle 6 Data Files

General Outline

* Mostly Omitted or Restricted-use items

FEMALE RESPONDENT FILE

A: Background and demographic information
B: Pregnancy and adoption-related information
C: Marital and relationship history; first sexual intercourse; recent partner history
D: Sterilizing operations and impaired fecundity
E: Contraceptive history and related information
F: Family planning and medical services
G: Birth desires and intentions
H: Infertility services and reproductive health
I: More background, more demographic information, and attitude questions
*J: Audio-CASI: pregnancy reporting; drug use; STD/HIV-risk behaviors; nonvoluntary intercourse; income

Recodes (created variables) and imputation flags for Sections A-J
(including key recodes describing pregnancies)
Weights & related variables

FEMALE PREGNANCY (INTERVAL) FILE

B: pregnancy outcomes and dates, prenatal care, sources of payment for delivery, maternity leave, breast-feeding
E: contraceptive use in the pregnancy interval and wantedness of the pregnancy

Recodes and imputation flags for Sections B&E
Selected respondent file variables (e.g., race/ethnicity, age)
Weights and related variables

MALE RESPONDENT FILE

A: Background and demographic information
B: Sex education, vasectomy and infertility, sexual intercourse, number of sexual partners
C: Current wife or cohabiting partner: date of marriage; children; contraception with her
D: Recent sexual partner(s) (up to three): key dates, children; contraception with her; 1st partner
E: Former wives and first premarital cohabiting partner: key dates, children; contraception with each
F: Other biological and adopted children; other pregnancies
G: Fathering: Activities with children he (a) lives with (b) does not live with
H: Desires and intentions for future children
I: Health conditions, access to health care, and receipt of health services
J: More background, more demographic information, and attitude questions
*K: Audio-CASI: pregnancy reporting; drug use; STD/HIV-risk behaviors; nonvoluntary intercourse; income

Recodes and imputation flags for Sections A-K
Weights & related variables

Detailed Outline for Each of the Three Data Files

* Restricted-use data

Abbreviations: R= Respondent
H/P= Husband or cohabiting partner
W/P= Wife or cohabiting partner
DOB= Date of birth (recorded in "century months")

FEMALE RESPONDENT FILE: (One record per Respondent)

Questionnaire Items: Sections A-J

Section A:

Demographic characteristics: age, DOB, marital/cohabiting status, race & Hispanic origin
Household (HH) roster: Number of HH members, relationship of man in HH, location of H/P if not in HH
*HH roster – age, race, and sex of member; relationship of member to R
Life History Calendar explanation
Education: grade currently attending; degrees; highest grade completed; date of hs graduation
Childhood background: Always lived with both parents (during childhood) or not
Whether parents were married at R's birth
Living situation at age 14
Mother's education, work, age at first birth, children ever born
Father's education

Section B:

Menarche
Current pregnancy status
Number of pregnancies
Relinquishment: number of children placed for adoption by R
Care of other children not born to R
Adoption plans (current and past), adoption preferences, & reasons for stopping pursuit

Section C:

Number of marriages and details on each husband
If currently cohabiting: Details on current cohabiting partner
Number of other cohabiting partners and further details on these partners
Ever had sexual intercourse *(asked only if never pregnant, never married, and never cohabited)*
IF NO: main reason why R has not had intercourse up to now
IF YES: Age at first intercourse
Date of first intercourse

Details on first sexual partner: (if not already discussed as husband or cohabiting partner)
Verification questions to get date & age of first intercourse after menarche
Sex education/communication items (only asked for teens), including timing relative to first sex
Number of sexual partners: In lifetime
In last 12 months
Details on all sexual partners in last 12 months: More details if he is a current partner
Clarification of R's *last sexual partner* (if R had more than 1 partner in last month of sexual activity)
2 questions on R's *last sexual partner* (if R had no sexual partners in last 12 months)

Section D:

Sterilizing operations (details on each operation R has had and R's H/P has had)
Desire for sterilization reversal (only for tubal ligations & vasectomies)
Sterilizing operations among former husbands and cohabiting partners (date; if reversed, date of reversal)
Non-surgical sterility & impaired fecundity

Section E:

Ever-use of individual birth control methods for any reason
Ever discontinue (dissatisfied); reasons for dissatisfaction with selected methods
First method *ever used* & details
Method use at first sexual intercourse
Months of non intercourse for past 4 or 5 years or since first intercourse (later of the 2 dates)
Contraceptive method history *(month by month, for past 3 or 4 years or since first method use)*
Method use at 1st and last sex with up to 3 partners in the past 12 months
Current method use/nonuse questions
Recent pill use reasons and brand/type
Consistency of condom use (including frequency of sex in past 4 weeks)

Section F:

Birth control Services include: birth control method; check-up or medical test related to using birth control; counseling about birth control; counseling about getting sterilized; sterilization; emergency contraception; information about emergency contraception.
Medical services include: pregnancy test; abortion; Pap smear; pelvic exam; prenatal care; post-pregnancy care; testing or treatment for sexually transmitted disease (STD)

Birth control and medical services in past 12 months
Provider & payment information for each visit and whether regular source of medical care
More information about the provider if they reported a clinic
Receipt of free condom, foam, or oral contraceptives from a clinic
First visit for birth control services: date, what services, & provider
Ever visited clinic, for those who did not report visiting a clinic in the last 12 months

Section G:

Wanting a/nother baby (R & H/P)
Intending a/nother baby *(joint or individual as appropriate)* & number intended

Section H:

Medical help to get pregnant
Medical help to prevent miscarriage
Specific infertility diagnoses received (if ever pursued medical help)
Vaginal douching: frequency in last 12 months; timing (post-coital vs. other times)
Health problems related to childbearing
    Pelvic Inflammatory Disease (PID): selected details
    Ever had: diabetes (gestational, nongestational), ovarian cyst, fibroid tumors or myomas in uterus, endometriosis, problems with ovulation or menstruation
2-question series on disability
HIV testing: Ever had
    Selected details about most recent test *(including new items on testing during pregnancy, if recently pregnant, and knowledge of preventive treatment for perinatal transmission)*

Section I:

Health insurance coverage in last 12 months
Residence April 1, 2000 (general)
Whether born outside U.S.; year came to U.S. to stay *(if born outside U.S.)*
Rent/own/payment for current residence
Religion: religion raised; frequency of attendance at age 14; current religion & frequency
Work: date of first full-time work (lasting 6+ months), ever not work full-time (lasting 6+ months)
    Number of months R worked in past 12 months
Current work status, # of jobs, full- or part-time
H/P's current work status # of jobs, full- or part-time
Child care use (type) in past 4 weeks for children under 13

Attitudes: Premarital sex, reaction to pregnancy *(teens)*, parenthood, marriage, divorce, cohabitation, gender roles, condom use

Section J:

*General health; height & weight
*Pregnancy reporting -- numbers ending in live birth, abortion, or other outcomes; total number
*Substance use: cigarettes, alcohol, marijuana, cocaine, crack, IV drugs
*Sex with males:
    Specific sexual behaviors that R may have engaged in
    Condom use at last occurrence of vaginal, oral, or anal sex
    Condom use at last occurrence of any type of sex & reasons why
    Nonvoluntary sex series *(asked only for 18-44)*
    Series of HIV/STD risk behaviors, including numbers of male partners
*Sex with females, including numbers of female partners
*Sexual health & HIV/STDs, including sexual attraction and orientation
Family income, sources of income, and public assistance in last full calendar year (2001)

Recodes and associated imputation flags: Female Sections A-J

(See Appendix 2 for further details)

A: Age; formal marital status; race/ethnicity; education (highest level, degrees); Hispanic origin, race; household composition; intact family status during childhood; parental living arrangements at age 14; R's mother's education; age of mother at her first birth

B: Current pregnancy status; total number of pregnancies; outcome-specific pregnancy counters; parity and number of births in last 5 years; number of children born out of wedlock and in cohabiting unions; descriptors of 1st birth; selected pregnancy-based recodes for pregnancies 1-19 (pregnancy outcome, end date, end year, conception date; R's age and formal marital status at conception & outcome; R's informal marital status at outcome) *(included on respondent file for data user's convenience)*

C: Marriage start/end dates and mode of dissolution; cohabitation experience relative to 1st marriage; duration and outcome of 1st cohabitation; 1st sex dates and ages (1st sex ever and 1st sex after menarche); 1st partner characteristics; intervals between 1st sex and other key dates; whether R had sex only once; numbers of sexual partners in lifetime and in last 12 months; date of last sex

D: Sterilization, fecundity, and infertility status

E: Current contraceptive status, source of method used in month prior to interview, 1st method use (type & date), method used at 1st sex, nonintercourse (last 12 months & last 36 months), ever-use of selected methods; recent condom use; wantedness recodes for pregnancies 1-19 (both definitions for R and partner) *(included on respondent file for data user's convenience)*; number of wanted pregnancies in past 5 years

F: Type of clinic used for family planning services in last 12 months; Type of clinic used for medical services in last 12 months; Title X clinic used for family planning service was regular place for care; Title X clinic used for medical service was regular place for care

G: Intentions for additional births; Central number of additional births expected

H: Infertility services and diagnoses received; PID experience; HIV testing

I: Current health insurance coverage; metropolitan residence; religion; labor force status

J: Poverty level of household income; total household income; receipt of public assistance in last year

Weights and related variables

Date of interview and time stamps

FEMALE PREGNANCY (INTERVAL) FILE
(1 record per pregnancy; No respondent had more than 19 pregnancies)

Questionnaire Items: Sections B and E

Section B

Pregnancy outcome *(none with more than 2 mentions)*
If current pregnancy – gestational length
If completed pregnancy:
- End date
- Gestational length
- Age of father when pregnancy ended/child was born
- *If pregnancy ended in or after Jan '97 and was not an induced abortion:*
  - Smoking during pregnancy
  - When R learned she was pregnant
  - Timing of first prenatal care visit
- *If pregnancy ended in live birth:*
  - Number of babies born alive
  - Payment for delivery *(if Jan '97 or later)*
  - Maternity leave (duration, duration paid) *(if Jan '97 or later)*
  - *For each of up to 3 babies:*
    - Sex
    - Birth weight *(if DK/RF weight, was child low birth-weight)*
    - Child's current living arrangement & date stopped living with R
    - Breast-feeding (ever breast-fed; child's age when supplemented and when stopped breast-feeding altogether)

Section E

Conditions surrounding R becoming pregnant
- method use in pregnancy interval
- whether all methods were stopped prior to that pregnancy
- whether absence of method-use was because R desired pregnancy
- method(s) R was using when she became pregnant that time
- wanted pregnancy at any time in future
- follow-up confirmation question for those who said never wanted pregnancy *(under age 20 only)*
- timing of pregnancy
- for sooner than wanted, how much too soon
- wanted pregnancy with that partner
- 10-point scale of happiness about that pregnancy
- father of preg: wanted pregnancy at any time in future
- father of preg: timing of pregnancy
- whether living with father of the preg. at the time of conception and pregnancy end/birth
- whether and when told father of preg about the pregnancy
- *If pregnancy ended in Jan '99 or later (and for current pregnancies):*
- 10-point scale of how hard trying not to get pregnant
- 10-point scale of how much wanted to avoid getting pregnant
- reasons for becoming pregnant with unwanted/mistimed pregnancy

Recodes and associated imputation flags: Sections B&E

B: Pregnancy outcome, gestational length, end date, end year, conception date; R's age at outcome and conception, formal marital status at outcome and conception; informal marital status at outcome; weeks pregnant when R learned of pregnancy and when R began prenatal care; payment for delivery; low birthweight for 1st baby; duration of breast-feeding; maternity leave

E: Wantedness of pregnancy, both Cycle 4 version and Cycle 5 version, for R and partner

Respondent File Variables (included on pregnancy file for data user's convenience)

Age at interview and at screener
Formal marital status at interview
Informal marital status at interview
Education (highest level, number of years)
Race; Hispanic or Spanish origin; Race & Hispanic origin combined
Whether R is currently pregnant at interview
Number of pregnancies overall
Number of live-born children overall (parity)
Health insurance coverage at interview
Received public assistance in past year
Poverty level income
Labor force status
Religious affiliation
Metropolitan residence
Born outside U.S.? If no, year R came to U.S. to stay

Weights and related variables
Date of interview

MALE RESPONDENT FILE (One record per Respondent)

Questionnaire Items: Sections A-K

Section A:

Demographic characteristics: age, DOB, marital/cohabiting status, race & Hispanic origin
Household (HH) roster: Number of HH members, relationship of woman in HH, location of W/P if not in HH
*HH roster – age, race, and sex of member; relationship of member to R
Education: grade currently attending; degrees; highest grade completed; date of hs graduation
Childhood background: Always lived with both parents (during childhood) or not
Whether parents married at R's birth
Female and male parent/parent-figure at age 14
Mother's education, work, age at first birth, children ever born
Father's education
Number of women married to and number of other women he lived with

Section B:

Sex education series *(teens only)*
Vasectomy: date of operation and reversal *(if any)*; if within last 5 years, also ask place
Infertility (self-assessed)
Ever had sex *(asked only if never married or cohabited)*
Main reason why R has not yet had intercourse *(for those who never had sex)*
Number of biological children
Number of sexual partners in lifetime and in last 12 months
Recent condom use consistency
Enumeration of recent (last 12 months) sex partners (up to 3)
    Establishment of whether any of these 3 partners was R's first sexual partner

Section C: Current wife or cohabiting partner

Key dates of marriage, cohabitation with current wife or cohabiting partner
Background/demographic information on current wife or cohabiting partner
First sex: date, method use, relationship at first sex *(if she is first partner ever)*
Her sterilizing operations and infertility
Last sex: date, method use
Method use in last 12 months: method most used, % of times used condom, fraction of time used no methods
Biological children with her, including conditions surrounding pregnancy if born in last 5 years
Current pregnancy with her, including conditions surrounding pregnancy
Information about other children she had at start of marriage or cohabitation who R adopted
Information about other non-biological children they raised/adopted together *(if any)*

Section D: Up to 3 most recent partners in last 12 months or Last Partner ever; 1st partner ever

Key dates of marriage, cohabitation, separation, divorce, widowhood with each
Last sex: date, method use
Background information on partner
First sex: date, method use, relationship at first sex *(if she is first partner ever)*
Method use in last 12 months: method most used, % of times used condom, fraction of time used no methods
Similar information on children and current pregnancies as in Section C
Data on first sexual partner ever: age, date, method use, relationship at first sex

Section E: Former wives and 1st premarital cohabiting partner

Key dates of marriage, cohabitation, divorce, widowhood
Background information on each former wife and the first cohabiting partner
Similar information on children as in Section C

Section F:

Number *(if any)* of additional biological children not already talked about
    Information about all other biological children (as above in Section C)
    Age of each mother of these children
Number *(if any)* of additional adopted children not already talked about
    Information about all other adopted children (as above in Section C)
Number of pregnancies that ended in miscarriage, stillbirth, abortion
Number of sexual partners in lifetime and in last 12 months

(if 7 or more partners were reported in Section B)

Section G:

Involvement with coresidential children under age 19:
- Activities in the last 12 months (all, respondents with children 5-18)
- Activities with their children in the last 4 weeks, based on age of children (<5, 5-18)
  - R with children aged 5-18: Include help with homework, talk, taking to activities, eating meals together
  - R with children under 5: Include feeding, bathing, playing, and reading
- Self-rating as father (for total group of coresidential children)

Involvement with non-coresidential biological or adopted child under age 19:
- How often visit in the last 12 months; how satisfied with number of visits
- ctivities with children (same as series for coresidential children)
- Self-rating as father (for group of noncoresidential children)
- Financial support of noncoresidentail children: any contribution in last 12 months; frequency & amount; contribution result of child support agreement

Section H:

Wanting a/nother baby (R)
Intending a/nother baby *(joint or individual as appropriate)* & number intended

Section I:

Usual source *(if any)* for medical care
Health insurance coverage in last 12 months
Whether ever accompanied female to family planning clinic and recency of last visit *(if R < 25 years old)*
His own receipt of services from family planning clinic: More details if within last 12 mos
2-item series on disability
Specific health services received in last 12 months: number of visits, provider, payment
Medical help for infertility & Infertility diagnoses received
HIV testing: Ever tested and selected details about most recent test

Section J:

Residence April 1, 2000 (general)
Whether born outside U.S.; year came to U.S. to stay *(if born outside U.S.)*
Rent/own/payment for current residence
Religion: religion raised; frequency of attendance at age 14; current religion & frequency
Ever served in military and starting/ending years of service
Work: date of first full-time work (lasting 6+ months), ever not work full-time (lasting 6+ months)
number of months R worked in past 12 months
Current work status, # of jobs, full- or part-time
W/P's current work status # of jobs, full- or part-time

Attitudes: Premarital sex, reaction to pregnancy *(teens)*, parenthood, marriage, divorce, cohabitation, gender roles, condom use

Section K:

*General health; height & weight
*Significant life events in last 12 months
*Substance use: alcohol, marijuana, cocaine, crack, IV drugs
*Pregnancy/abortion: Ever caused pregnancies/births/abortions; if so, how many
If R < 25: Ever *told* so; if so, what happened the last time
*Sex with females:
Specific sexual behaviors that R may have engaged in
Condom use at last occurrence of vaginal, oral, or anal sex
Condom use at last occurrence of any type of sex & reasons why
Nonvoluntary sex series *(asked only for 18-44)*
Series of HIV/STD risk behaviors, including numbers of female partners
*Sex with males:
Specific sexual behaviors that R may have engaged in
Condom use at last occurrence of oral or anal sex
Condom use at last occurrence of any type of sex
Nonvoluntary sex series *(asked only for 18-44)*
Series of HIV/STD risk behaviors, including numbers of male partners
*Condom use at last sex of any type (and if it was vaginal sex with a female, reasons why)
*Sexual health & STDs/HIV, including sexual attraction and orientation
Family income, sources of income, and public assistance in last full calendar year (2001)

Recodes and associated imputation flags: Male Sections A-K
(See Appendix 2 for further details)

A: Age; formal marital status; race/ethnicity; education (highest level, degrees); Hispanic origin, race; household composition; intact family status during childhood; parental living arrangements at age 14; R's mother's education; age of mother at her first birth; number of marriages; informal marital status

B-F: Date and age at 1st sex; whether R had sex only once; characteristics of 1st partner; method use at 1st sex; date of last sex and partner characteristics; method use at last sex, at last sex in last 3 months, and at last sex in last 12 months; numbers of partners in lifetime and in last 12 months; cohabitation experience relative to 1st marriage; duration and outcome of 1st cohabitation; dates of 1st marriage & dissolution; how 1st marriage ended; duration of 1st marriage; premarital cohabitation with 1st wife; date of 1st biological child's birth; whether 1st child was born premaritally; formal marital status at 1st child's birth; numbers of children born out of wedlock and in cohabiting unions; number of out-of-wedlock children with paternity establishment; summary counts of pregnancies fathered, by outcome; wantedness of biological children; number of unintended births in last 5 years

G: Type of children aged 18 or younger that R has; Type of children under 5 that R has; Type of children 5-18 that R has; Number of coresidential children under 18; Number of noncoresidential children under 18; contribution of child support in the last 12 months

H: Intentions for additional births; Central number of additional births expected

I: Current health insurance coverage; ever used infertility service, ever had HIV test

J: Metropolitan residence; religion; labor force status

K: Poverty level of household income; total household income; receipt of public assistance in last year

Weights and related variables

Date of interview and time stamps

Combining Data from Female Respondent and Pregnancy Files Using SAS

As mentioned in the section called "Organization and Use of the Data File," selected pregnancy (interval) variables have been placed on the female respondent file, and selected female respondent variables have been placed on the pregnancy (interval) file, but users may need to merge in additional variables for their analyses. Below are 2 examples of SAS programs that will yield 1) a pregnancy-based file and 2) a respondent-based file. The user must tailor these examples to their own computing environments (e.g., adding their own file definition statements).

Example 1: Adding Respondent Variables to a Pregnancy (Interval) Based File

This template program will yield a sasfile with 13,593 records, assuming that the user does not subset observations from the pregnancy file. The respondent-based variables that are not already on the pregnancy file will be added to EACH pregnancy record with the same CASEID (case identification number).

```
DATA RESPOND;
        INFILE IN1;
        INPUT CASEID $ 1-12
                :       (insert other variables desired from respondent file)
                :
        ;
DATA PREG;
        INFILE IN2;
        INPUT CASEID $ 1-12
                :       (insert other variables desired from pregnancy file)
                :
        ;
DATA PREGFILE;
        MERGE RESPOND PREG (IN=A);
        BY CASEID;
        IF A;
```

Example 2: Adding Pregnancy Variables to a Respondent Based File

This template program will extract the most recent live birth for each respondent from the pregnancy (interval) file and merge with selected variables from the respondent file. The resulting sasfile will yield a respondent-based sasfile with less than 7,643 records, because only those respondents who have ever had a live birth will be included. This program may be helpful if the user wishes to examine, for example, breastfeeding and maternity leave for the most recent birth. For this program, the following pregnancy variables are needed:

CASEID = Case identification number
OUTCOME = Outcome of pregnancy (=1 if live birth)
PREGORDR = Pregnancy order or number

```
DATA RESPOND;
        INFILE IN1;
        INPUT CASEID $ 1-12
                :       (insert other variables desired from respondent file)
                :
        ;
DATA PREG;
        INFILE IN2;
        INPUT CASEID $ 1-12 PREGORDR 13-14 OUTCOME 277
                :       (insert other variables desired from pregnancy file)
                :
        ;
        IF OUTCOME=1;     /* keep only live births */

PROC SORT; BY CASEID PREGORDR;      /* sort PREGORDR within CASEID */

DATA LASTPREG;
        SET PREG; BY CASEID;
        IF LAST.CASEID THEN OUTPUT; /* keep only the last birth for each R */

DATA LASTBRTH;
        MERGE RESPOND LASTPREG (IN=A);
        BY CASEID;
        IF A;
```

Key to the File Indexes

These file indexes (located in Appendix 1) are a listing of each data item on the Female Respondent, Female Pregnancy (Interval), and Male Files, and includes: the variable name, data file column location, a short item description including the question number if applicable, and an indication of the variable type.

Variable name:

Corresponds exactly or approximately to the question name in the CAPI Reference Questionnaire (CRQ) (see section on "Questionnaires"). In some cases, one question in the questionnaire yields several variables in the data file. This occurs for two reasons:
(1) the question may have been repeated for multiple occurrences of an event or multiple individuals (for example, sexual partners in the last 12 months in the female questionnaire, biological children fathered with a given wife or partner in the male questionnaire); or
(2) the question may be a "code all that apply" item (for example, forms of payment for family planning services, infertility services ever received, contraceptive methods used in a given month).
In these instances, the *variable name* appearing in the index corresponds to the first several (4-7) characters of the *question name* appearing in the CRQ, followed by numbers. Note that the numbers appearing on the variable names in the index do not always correspond to the actual number of the occurrence of the event or the order of the individual (for example, PAYRSTER7 on the female respondent file indicates the 2nd form of payment reported for a hysterectomy). Therefore, it is critical to check the full item description, which will always contain the correct number or iteration.

Item description and variable type:

For original data items asked during the interview, the CRQ question number appears in the item description. The female survey was organized into 10 sections roughly corresponding to substantive topics and the male survey was organized into 11 sections (see "outline of contents of the data files"). **The first letter of the question number indicates the section of the questionnaire to which the question belongs.**
Item descriptions without question numbers are one of the following:
(1) computed variables, created during the interview for the purposes of the computer-assisted survey program;
(2) recodes, created from 1 or more original data items, after data collection was completed (see section "recodes"); or
(3) other miscellaneous variables, such as "intermediate variables" defined for creating recodes and "weights and weight-related variables."
The variable type of each of these variables is indicated in the file indexes as well.

Key to Codebook Documentation

For each item, the codebook documentation includes a variable name, numbers indicating the beginning and ending column locations on the data file, the question text or description for the item, the "applicable specification" (i.e., universe statement), the numerical values into which responses were coded with detailed value labels, and unweighted frequencies.

Variable name:

For original data items that appeared in the survey, the variable name corresponds exactly or approximately to the question name that appears in the CAPI Reference Questionnaire, which is the readable version of the interview program. (See "Key to File Indexes" for further details on variable names.)

Question text:

For original data items that appeared in the survey, the question text corresponds to the actual questionnaire item text and is preceded by the question number. Where appropriate, question wording variants are presented, but given the often complex tailoring of some question wording, the user may wish to consult the CAPI Reference Questionnaire for the precise wording and question routing (see section on "Questionnaires").

For variables that were computed as part of the Blaise survey instrument, the question text ("description") corresponds in large part to the item description shown in the File Indexes, and concludes with "(COMPUTED)" to indicate the variable type.

For recodes and other variables not originally in the survey, the question text ("description") corresponds to the item description that appears in the file index (in Appendix 1) and/or in the recode specifications (Appendix 2). Question text for these variables ends with "(RECODE)" or "(INTERMEDIATE VARIABLE)" to indicate the variable type.

Applicable specifications (universe statements):

For each item, beneath the question text is a statement specifying the universe of respondents for whom the item is applicable. If the item was applicable for all respondents, it was described as such. If the item was **not** applicable for all respondents, you will find an "applicable specification," a description of the conditions under which the item was applicable. For further information, see section on "Universe Statements."

Response categories and unweighted frequencies:

For categorical variables and several continuous variables, the documentation lists all possible values with descriptive value labels. Variables that are not applicable for all respondents include the number of "inapplicable" cases. Most century month (date) and continuous variables have been collapsed into meaningful, yet manageable, groups. Every question in the NSFG interview had a "don't know" and "refuse" option. These are only presented among the response categories if one or more cases gave such a response to the question. In addition, some variables have one or more cases with "not ascertained" values, and this category is shown where it occurs (see section on "Missing Data" for more information).