SAFER • HEALTHIER • PEOPLE™

CDC™

Monitoring the Nation's Health

Vital and Health Statistics

Series 2, Number 142

July 2006

# National Survey of Family Growth, Cycle 6: Sample Design, Weighting, Imputation, and Variance Estimation

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

**Copyright information**

All material appearing in this report is in the public domain and may be reproduced or copied without permission; citation as to source, however, is appreciated.

---

**Suggested citation**

Lepkowski JM, Mosher WD, Davis KE, et al. National Survey of Family Growth, Cycle 6: Sample design, weighting, imputation, and variance estimation. National Center for Health Statistics. Vital Health Stat 2(142). 2006.

---

**Library of Congress Cataloging-in-Publication Data**

National Survey of Family Growth, cycle 6 : sample design, weighting, imputation, and variance estimation.
p. ; cm.— (Vital and health statistics. Series 2, Data evaluation and methods research ; no. 142) (DHHS publication ; no. (PHS) 2006–1342)
"July 2006."
Includes bibliographical references.
ISBN 0–8406–0608–7
1. Family size—Statistical methods. 2. Fertility, Human—Statistical methods. 3. Family life surveys. 4. Health surveys. I. National Center for Health Statistics (U.S.) II. Series. III. Series: DHHS publication ; no. (PHS) 2006–1342
[DNLM: 1. National Survey of Family Growth (U.S.) 2. Research Design—United States. 3. Data Collection—methods—United States. 4. Family Characteristics—United States. 5. Parenting—United States. 6. Reproductive Behavior—United States. W2 A N148vb no.142 2006]
RA409.N38 no. 142
[HQ762]
614.4'2—dc22 2006014336

---

For sale by the U.S. Government Printing Office
Superintendent of Documents
Mail Stop: SSOP
Washington, DC 20402-9328
Printed on acid-free paper.

# Vital and Health Statistics

Series 2, Number 142

# National Survey of Family Growth, Cycle 6: Sample Design, Weighting, Imputation, and Variance Estimation

## Data Evaluation and Methods Research

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

Hyattsville, Maryland
July 2006
DHHS Publication No. (PHS) 2005-1342

## National Center for Health Statistics

Edward J. Sondik, Ph.D., *Director*

Jennifer H. Madans, Ph.D., *Acting Co-Deputy Director*

Michael H. Sadagursky, *Acting Co-Deputy Director*

Jennifer H. Madans, Ph.D., *Associate Director for Science*

Edward L. Hunter, *Associate Director for Planning, Budget, and Legislation*

Michael H. Sadagursky, *Associate Director for Management and Operations*

Lawrence H. Cox, Ph.D., *Associate Director for Research and Methodology*

Margot A. Palmer, *Director for Information Technology*

Margot A. Palmer, *Acting Director for Information Services*

Linda T. Bilheimer, Ph.D., *Associate Director for Analysis Epidemiology, and Health Promotion*

Charles J. Rothwell, *Director for Vital Statistics*

Jane E. Sisk, Ph.D., *Director for Health Care Statistics*

Jane F. Gentleman, Ph.D., *Director for Health Interview Statistics*

Clifford L. Johnson, *Director for Health and Nutrition Examination Surveys*

### Division of Vital Statistics

Charles J. Rothwell, M.S., *Director*

James A. Weed, Ph.D., *Deputy Director*

Stephanie J. Ventura, M.A., *Chief, Reproductive Statistics Branch*

Nicholas F. Pace, *Systems Programming and Statistical Resources Branch*

# Acknowledgments

This report describes the sample design, associated estimation procedures, and variance estimation methods used in Cycle 6 of the National Survey of Family Growth. The survey was designed and conducted by the National Center for Health Statistics (NCHS), in cooperation with the survey contractor, the Institute for Social Research (ISR), University of Michigan, Ann Arbor, Michigan. The sampling plan was developed by Robert M. Groves, Steven G. Heeringa, and James M. Lepkowski, Institute for Social Research (ISR), in consultation with William Mosher and Karen Davis of NCHS. The information in this report is based on survey design documentation prepared by staff of ISR and on internal NCHS memoranda.

Cycle 6 of the National Survey of Family Growth (NSFG) was jointly planned and funded by the following programs and agencies of the U.S. Department of Health and Human Services:

- The National Institute of Child Health and Human Development
- The Office of Population Affairs
- The CDC's National Center for Health Statistics (CDC/NCHS)
- The CDC's National Center for HIV, STD, and TB Prevention
- The CDC's Division of Reproductive Health
- The CDC's Office of Women's Health
- The Office of Planning, Research, and Evaluation of the Administration for Children and Families (ACF)
- The Children's Bureau of the ACF
- The Office of the Assistant Secretary for Planning and Evaluation (OASPE)

NCHS gratefully acknowledges the contributions of these programs and agencies, and all others who assisted in designing and carrying out the NSFG. Steven Heeringa and Terry Adams of the University of Michigan reviewed a preliminary draft of this report and provided useful comments. Two independent referees also reviewed this report: Christopher L. Moriarity, formerly of the Government Accountability Office (GAO), and Steven R. Machlin of the Agency for Healthcare Research and Quality. Their suggestions have been incorporated into the report; any remaining errors are the responsibility of the authors and not the reviewers.

This report was prepared under the general direction of Charles J. Rothwell, Director of the Division of Vital Statistics (DVS), James A. Weed, Deputy Director of DVS, and Stephanie J. Ventura, Chief of the Reproductive Statistics Branch (RSB). The report was edited by Gail V. Johnson of the Information Design and Publishing Staff, Office of Information Services, NCHS; typesetting was done by Zung Le and graphics were produced by Jarmila G. Ogburn, CoCHIS/NCHM/ Division of Creative Services.

# Contents


Acknowledgments . . . iii

Abstract . . . 1

Executive Summary . . . 1
- Sample Design . . . 2
- Responsive Design Features . . . 2
- Sampling Weights . . . 2
- Item Imputation . . . 3
- Variance Estimation . . . 3
- Conclusion . . . 4

Background . . . 4

Design Specifications . . . 4

Sample Design . . . 5
- Summary of Sample Design . . . 5
- Sample Size Requirements . . . 8
- First-Stage Selection of Primary Sampling Units . . . 9
- Second-Stage Selection of Segments . . . 11
- Third-Stage Selection of Housing Units . . . 12
- Fourth-Stage Selection of Sample Persons . . . 13
- Sample Release . . . 14

Sampling in Three Phases . . . 15

Sample Characteristics . . . 16
- Response Rate Estimation . . . 16
- Nonresponse Bias and Phase 3 Sampling . . . 19
- Cost-Efficiency and Response Rates in the Phase 3 Sample . . . 20
- Responsive Design . . . 22
- Sample Size . . . 24

Weighting Procedures . . . 25
- Summary . . . 25
- Inverse Probability Selection Weighting . . . 25
- Nonresponse Adjustment . . . 28
- Poststratification . . . 31
- Weight Trimming . . . 31
- Final Weight Summary . . . 32
- Estimating Equation for Totals . . . 33

Imputation Procedures . . . 34
- Background on Imputation for Missing Values . . . 34

Sequential Regression Imputation . . . 35

Multiple Imputation Using Sequential Regression . . . 36
- Regression Imputation in NSFG Cycle 6 . . . 37

Variance Estimation . . . 37
- Background . . . 37
- Summary of Variance Estimation Principles . . . 37

Variance Estimation Software . . . 38
Application to the NSFG . . . 39

References . . . 44

Appendix I . . . 46
Definition of Terms . . . 46

Appendix II . . . 49
Variables Imputed and Types of Imputation for Female, Male, and Pregnancy Files . . . 49

Appendix III . . . 55
Distribution of Female, Male, and Pregnancy Records, by Stratum and Sampling Error Computing Units in the National Survey of Family Growth, Cycle 6 . . . 55

Appendix IV . . . 63
Variance Estimates Using SAS, SUDAAN, and STATA, for a Percentage . . . 63

Appendix V . . . 68
Variance Estimates Using SAS, SUDAAN, and STATA, for a Mean . . . 68

Appendix VI . . . 73
Variance Estimates Using SAS, SUDAAN, and STATA, for a Regression . . . 73

Appendix VII . . . 76
Variance Estimates SAS, SUDAAN, and STATA for a Logistic Regression . . . 76

Appendix VIII . . . 80
Median Design Effect for Female, Male, and Pregnancy File Means and Proportions . . . 80

## Figures

1. History of the National Survey of Family Growth . . . 5
2. National Survey of Family Growth Cycle 6 selection of primary sampling units and segments . . . 6
3. Selection steps in sample implementation for Cycle 6 of the National Survey of Family Growth . . . 7
4. Within-household screener and household roster . . . 8
5. Within-household measures of size and illustration of within-household selection . . . 9
6. Response rates for Phases 1 and 2 and overall for 18 subpopulations . . . 20
7. Response rates for Phases 1 and 2 and overall (reflecting subsampling of nonrespondents) by age group; National Survey of Family Growth Cycle 6, 2002 . . . 20
8. Estimated proportion of females who have ever had sexual intercourse in Cycle 6 of the National Survey of Family Growth: Cumulative estimate; estimate by call number of interview, and estimate from Phase 3 interviewing . . . 21
9. Cumulative ratio of interviewer hours to number of main interviews completed by day of the data-collection period: National Survey of Family Growth Cycle 6, 2002 . . . 24
10. SAS 9.1 program code for illustration 1 . . . 40
11. SAS 9.1 output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002 . . . 41
12. SUDAAN program code for illustration 1 . . . 42
13. SUDAAN 8.0.2 output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002 . . . 43
14. STATA 8.0 program code for illustration 1 . . . 44
15. STATA output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002 . . . 44
16. Wesvar table specification window for illustration 1 for estimation of the percent of women using the oral ontraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002 . . . 45

## Text Tables

A. Allocation of sample segments and sampling rates within domains: National Survey of Family Growth, Cycle 6 . . . 12
B. Expected and observed eligibility rates by sex, age, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 14
C. Number of Phases 1 and 2 segments by number of active incomplete cases in the segment at the end of Phase 2, and the estimated total propensity of active cases to complete an interview on the next visit: National Survey of Family Growth, Cycle 6 . . . 16

D. Phase 3 segment sampling fractions, by number of active cases in the segment, and the estimated total propensity of active cases in the segment to complete the interview on the next visit: National Survey of Family Growth, Cycle 6 . . . 16
E. Response rates, weighted only for subsampling in Phase 3, but not for other probabilities of selection, using two estimates (0.64 and 0.58) of the proportion of nonscreened households that had an eligible respondent: National Survey of Family Growth, Cycle 6 . . . 18
F. Fully weighted response rates, adjusted for all unequal probabilities of selection, using two estimates (0.64 and 0.58) of the proportion of nonscreened households that had an eligible respondent: National Survey of Family Growth, Cycle 6 . . . 19
G. Mean number of calls per active screener case by number of active cases and total propensity strata in Phase 3: National Survey of Family Growth, Cycle 6 . . . 21
H. Mean number of calls per active main interview case by number of active cases and estimated total propensity of cases in the segment to respond on the next visit: Phase 3 of National Survey of Family Growth, Cycle 6 . . . 22
J. Mean screener response rates for segments in Phase 3 by number of active cases and total propensity strata: National Survey of Family Growth, Cycle 6 . . . 22
K. Mean main interview response rates for segments during Phase 3 by number of active cases at the beginning of Phase 3, and the total propensity of cases in the segment to respond on the next visit: National Survey of Family Growth, Cycle 6 . . . 23
L. Mean ratio of number of screener interviews to number of total calls on screener cases by number of active cases in the segment, and by the estimated total propensity of cases in the segment to respond on the next interviewer visit: National Survey of Family Growth, Cycle 6 . . . 23
M. Mean ratios of number of main interviews to number of total calls on screener cases by number of active cases and total propensity strata: National Survey of Family Growth, Cycle 6 . . . 23
N. Number of completed interviews by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 25
O. Mean, minimum, and maximum untrimmed base weight, and potential increase in variance due to weighting (1+*L*), by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 27
P. Logistic Regression Model Predicting Response to the Screener, with 18 predictors (Phases 1–3): National Survey of Family Growth, Cycle 6 . . . 29
Q. Nine logistic regression models predicting response to the main interview (Phases 1–3): National Survey of Family Growth, Cycle 6 . . . 32
R. Mean, minimum, and maximum combined nonresponse adjustments, and potential increase in variance due to weighting, by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 33
S. Poststratification adjustment factors for 36 groups, by age, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 34
T. Mean, minimum, and maximum final (trimmed) weight, and potential increase in variance due to weighting, by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6 . . . 35

## Appendix Tables

I. Imputed case counts in the female file . . . 49
II. Imputed case counts in the male file . . . 53
III. Imputed case counts in the pregnancy file . . . 54
IV. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the combined female and male file . . . 55
V. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the pregnancy file, NSFG Cycle 6 . . . 59
VI. Median design effects for 10 female respondent file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6 . . . 80
VII. Median design effects for six female pregnancy file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6 . . . 81
VIII. Median design effects for five male respondent file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6 . . . 82

## Abstract

### Objectives

Cycle 6 of the National Survey of Family Growth (NSFG) was conducted by the National Center for Health Statistics in 2002 and early 2003. This report describes how the sample was designed, shows response rates for various subgroups of men and women, describes how the sample weights were computed to make national estimates possible, shows how missing data were imputed for a limited set of key variables, and describes the proper ways to estimate sampling errors from the NSFG. The report includes both nontechnical summaries for readers who need only general information and more technical detail for readers who need an in-depth understanding of these topics.

### Methods

The NSFG Cycle 6 was based on an independent, national probability sample of men and women 15–44 years of age. It was the first NSFG ever to include a national sample of men 15–44 as well as a sample of women. Fieldwork was carried out by the University of Michigan's Institute for Social Research (ISR) under a contract with NCHS. In-person, face-to-face interviews were conducted by professional female interviewers using laptop computers. In all, 12,571 women and men—7,643 females and 4,928 males—were interviewed, the largest NSFG ever done.

### Results

Analysis of NSFG Cycle 6 data requires the use of sampling weights and estimation of sampling errors that accounts for the complex sample design and estimation features of the survey. Examples of how to use several available software packages that incorporate complex design features in estimation, such as SAS, SUDAAN, and STATA, are presented.



**Keywords:** *survey methodology • response rates • imputation • variance estimation • National Survey of Family Growth*


# National Survey of Family Growth, Cycle 6: Sample Design, Weighting, Imputation, and Variance Estimation


*By James M. Lepkowski, Ph.D., Institute for Social Research, University of Michigan; William D. Mosher, Ph.D., and Karen E. Davis, M.A., National Center for Health Statistics; and Robert M. Groves, Ph.D., John van Hoewyk, Ph.D., and Jennifer Willem, M.S., Institute for Social Research, University of Michigan*


## Executive Summary

Cycle 6 of the 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 15–44 years of age. This Series 2 report is the technical documentation of the procedures used to select the sample, develop the sampling weights, impute missing data, and estimate sampling errors. The report will be useful to at least two types of readers: first, data analysts who intend to use the NSFG and need to understand the sample design, weighting, imputation, and variance estimation; and second, those interested in survey methodology who may wish to compare their procedures to those used in the NSFG. Parts of this report contain a great deal of technical detail. For readers seeking a general understanding of the survey procedures, this section provides a summary of the procedures used.

The NSFG is designed and administered by the National Center for Health Statistics (NCHS), an agency of the U.S. Department of Health and Human Services, in response to Section 306 of the Public Health Service Act, which directs NCHS to "collect statistics on. . .family formation, growth, and dissolution," as well as "determinants of health" and "utilization of health care." Accordingly, the purpose of the survey is to produce reliable national statistics on:

- Factors affecting pregnancy—including sexual activity, contraceptive use, infertility, and sources of family planning services
- Factors affecting marriage, divorce, cohabitation, and adoption
- What women and men do to raise their children
- Their 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 living in households in the United States were interviewed between mid-March 2002 and the end of February 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; 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 and children fathered; fatherhood involvement measures; and demographic and economic characteristics. In a self-administered part of the interview, women and men were also asked questions on behaviors related to the risk of sexually transmitted diseases, including the human immunodeficiency virus (HIV).

This summary 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.

## Sample Design

A total of 12,571 men and women were selected using probability sampling methods from 121 primary sampling units (PSUs). A PSU is a metropolitan area, a county, or a group of adjacent counties. PSUs 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 (including persons living away from the household in a college dormitory, sorority, or fraternity) were listed. If more than one eligible person 15–44 years of age was living at the address, one person was randomly selected and asked to do an interview.

## Responsive Design Features

Surveys with high response rate goals and limited budgets such as Cycle 6 of the NSFG, need a way to stay informed on how much interviewer labor and money are being spent on data collection; what areas and interviewers are having good results and poor results; and what types of nonresponse are most prevalent in each area. The Institute for Social Research (ISR) used a system called SurveyTrak to provide that information each working day.

Because the NSFG budget for data collection was limited, and significant changes to the questionnaires and procedures had to be made between the pretest and the main study, it was important to determine how well main study interviewers could perform the tasks they had been trained to do. SurveyTrak allowed study staff to monitor how many hours of effort were required to obtain interviews, whether interviewers worked as many hours as expected, and response rates. Project staff used this and other information to manage study data collection to keep it within budget. The management process led to modification of study procedures throughout the field period. Three phases of fieldwork occurred.

*Phase 1 of fieldwork*—In March 2002, an initial group of interviewers was trained and started fieldwork in a subsample of areas. Monitoring through the SurveyTrak system made it possible to make better estimates of interviewer production and survey costs for the rest of the main study.

*Phase 2 of fieldwork*—The full study sample was released June 2002 through January 2003. Interviewers recorded characteristics of neighborhoods (sample segments) and sampled households. The contractor collected and analyzed these data during fieldwork and used the data to allocate interviewer labor during the study.

*Phase 3 of fieldwork*—SurveyTrak data were used most effectively in the last month of fieldwork—February 2003. Approval was obtained to use somewhat larger incentives to encourage participation. Prediction models based on SurveyTrak data were used to select segments that had the largest number of people who were most likely to complete the interview. Field resources in the last month of the survey were focused on the selected Phase 3 segments. A large number of interviews were obtained and contributed to higher weighted response rates across all three phases of data collection. The Phase 3 interviews also provided indirect evidence that nonresponse bias was reduced by the addition of Phase 3 interviews.

A more complete description of these design phases and the generation and use of these data are reported in the NSFG Series 1 report. The procedures to produce and use these data to manage fieldwork are called ''responsive design'' features and were developed to yield higher response rates in key demographic groups in the sample than would have occurred without it.

## Sampling Weights

A simple random sample in which response rates and coverage were the same in every subgroup would be a ''scale model'' of the population. However, many survey samples are not ''scale models'' in that sense. If a ''scale model'' of the population were selected, smaller groups in the population would have too few observations in the sample to provide adequate precision for characteristics of interest for those groups. As a result, survey samples often select groups at different rates deliberately to over-represent smaller groups in the sample. This allows analysts the opportunity to answer key survey questions for the total population and for those small but often important groups of the population.

For example, in the NSFG, non-Hispanic black men and women were chosen to account for 19.6 percent of all respondents in the sample, even though they are 12.9 percent of the population 15–44 years of age. In addition to black men and women, Hispanic men and women, and teenagers of all races were also sampled at higher rates than non-Hispanic white adults in this cycle of the NSFG. ''Sampling weights'' adjust for these different sampling rates, response rates, and coverage rates so that unbiased 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 three factors. The first factor is the inverse of the probability that the case was selected. For example, if the probability of selection is 1 in 6,000, then the initial sampling weight is 6,000. The second factor is an adjustment for nonresponse, 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, gender, race, and Hispanic origin, provided by the U.S. Census Bureau (and, for military personnel living in a household outside of a military facility, by the Defense Manpower Data Center). This process is called poststratification.

## 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 are 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 has several advantages. It makes the data more complete and easier to use, eliminating the sometimes confusing decreases in sample size when cases with missing values have to be dropped from an analysis. It also allows all of the collected data to be used in analysis. In an analysis involving several different variables, cases may be dropped because they are missing for one variable, even though there are nonmissing values for the other variables available for those cases. In effect, the analyst discards collected data by deleting cases with item missing values. Finally, discarding cases with missing values implicitly assigns a value to the missing items. Effectively, discarding cases assigns the average of the nonmissing values to each of the missing values. Imputation is a procedure that attempts to improve on this assignment by assigning a replacement value for each item missing a value that is a prediction from other variables and not just the average of the same variable from cases without missing values.

In Cycle 6 of the NSFG, there are thousands of variables in the data file. Of these many variables, 356 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 microdata 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. A total of 2.0 percent of the values of all recoded variables, in male, female, and pregnancy files combined, were missing and subsequently imputed. The program also performed range and consistency checks to help prevent logically impossible answers. The two imputation techniques used in Cycle 6 were:

- Logical imputation
- Regression imputation

Logical imputation involves having a substantive expert (usually at NCHS) look at the case, and by examining related variables, assign a value for a missing value that is essentially an educated guess of the true value. 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. Slightly over 90 percent of the imputed values were imputed by regression imputation.

Except when a reported value 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 all the 125 million men and women 15–44 years of age.) 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 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 this report, and further examples are shown on the Internet at: http://www.cdc.gov/nchs/about/major/nsfg/nsfgvar.htm.

## Conclusion

This report is designed in part to give researchers using the NSFG the knowledge they need to conduct their research in a statistically valid way. Because of the complex sample design of the NSFG, analysts should use the weights in analysis and use software that will compute ''design-based'' estimates of sampling errors. Failure to use the weights and accurate variance estimates may lead to biased or inaccurate findings and conclusions.

# Background

The National Survey of Family Growth (NSFG) is an ongoing series of sample surveys designed to provide current information about childbearing, contraception, and related aspects of maternal and child health for the United States. Established in 1971 by the National Center for Health Statistics, the NSFG is a periodic survey. The first cycle was conducted in 1973, and subsequently conducted in 1976, 1982, 1988, and 1995. Cycle 6 was conducted in 2002–03.

The NSFG target population has been expanded over time. For example, the target population for Cycles 1 and 2 was the civilian household population of women ages 15–44 years living in the 48 coterminous States and the District of Columbia. These women were currently or previously married, or never married mothers with children living with them at the time of interview (1,2).

The population for Cycle 3 included women of all marital statuses and added women living in group quarters. Thus, Cycle 3 included all women ages 15–44 years in the civilian noninstitutionalized population (3). Similarly, Cycles 4 and 5 targeted all women ages 15–44 years in the civilian noninstitutionalized population, but drew the NSFG sample from a larger NCHS survey, the National Health Interview Survey, and included Alaska and Hawaii in the sample (4–6).

The target population for NSFG Cycle 6 differs from previous cycles in two important ways. First, it includes both men and women 15–44 years of age in all 50 States and the District of Columbia. This is the first time men have been included in the NSFG. Second, the target population is restricted to the household population (including persons living away from home in college dormitories, sororities, or fraternities), but it does not eliminate all members of the military, as in past cycles. If a member of the military resided in housing outside a military base or installation, they were eligible to be included in the sample. Members of the military residing in military installations were not eligible. Cycle 6 thus includes for the first time men and military personnel residing in civilian housing.

In all cycles, data collection has been conducted in person. In Cycles 1–3, interviews lasted an average of 1 hour; in Cycle 4, the average interview length was 70 minutes; and in Cycle 5, 100 minutes (figure 1). Cycle 6 interviews were designed to last an average of 1 hour for males and 80 minutes for females. Across the six cycles, data are available on trends and changes in fertility, contraception, breastfeeding, use of family planning services, and maternal and child health.

The sample design and data collection for NSFG have been conducted by several different survey research organizations. For example, Cycle 1 was conducted by the National Opinion Research Corporation of the University of Chicago. Cycles 2, 3, and 4 were conducted by Westat, Inc. of Rockville, MD. Cycle 5 was conducted by the Research Triangle Institute of North Carolina, while the present cycle was conducted by the ISR at the University of Michigan. Descriptions of sample design and data collection procedures for Cycles 1 through 5 are available in reports produced by NCHS (1–6).

Cycle 6 obtained 12,571 completed interviews, 7,643 with eligible women and 4,928 with eligible men. Interviews were conducted between March 2002 and February 2003, with the midpoint of the survey in September 2002. This report describes the sample design developed to select the sample of men and women, procedures to make national estimates from the sample data, and methods used to estimate the precision of those estimated population parameters.

# Design Specifications

Cycle 6 employed a stratified multistage probability sample of households and eligible persons. The decision to use this kind of a design was based on the following specifications:

- The target population for Cycle 6 was to be the household population of men and women ages 15–44 years living in households in the 50 States and the District of Columbia.
- The sample was to consist of approximately 7,500 female and 4,800 male respondents from a probability sample of households. The sample was designed to yield at least 2,000 completed interviews with each of three groups: black persons 15–44 years of age, Hispanic persons 15–44 years of age, and teens 15–19 years of age.
- Screening interviews were to be conducted in each sample household, to determine eligibility and to select one person from each household that contained one or more eligible persons using a random selection procedure applied to the list of persons living in the household.
- Data collection was to be conducted by personal interview conducted in sample households. Interviews were expected to last an average of 1 hour for men and 80 minutes for women. No proxy reports or data were allowed.
- All interviewers were to be female because of the sensitivity of the content of the interview. That is, men and women were to be interviewed by a female interviewer.
- Data collection was to be completed using Computer-Assisted Personal Interviewing (CAPI). Survey instruments were to be programmed on laptop computers meeting specified requirements. Data collection was to be by personal

| Cycle | Year | Scope | Number of interviews | Source of sample | Over-samples | Average length in minutes | Incentive payment |
|---|---|---|---|---|---|---|---|
| 1 | 1973 | Ever-married women 15-44 | 9,797 | Independent 101 PSUs | Black women | 60 | No |
| 2 | 1976 | Ever-married women 15-44 | 8,611 | Independent 79 PSUs | Black women | 60 | No |
| 3 | 1982 | All women 15-44 | 7,969 | Independent 79 PSUs | Black women and teens | 60 | No |
| 4 | 1988 | All women 15-44 | 8,450 | NHIS 156 PSUs | Black women | 70 | No |
| 5 | 1995 | All women 15-44 | 10,847 | NHIS 198 PSUs | Black and Hispanic women | 100 | $20 |
| 6 | 2002 | Men 15-44 Women 15-44 | *12,571 | Independent 121 PSUs | Black and Hispanic persons and teens | M=60 F=85 | $40 |

*In 2002 National Survey of Family Growth surveyed 7,643 women and 4,928 men.
NOTE: PSU is primary sampling unit, NHIS is National Health Interview Survey.

**Figure 1. History of the National Survey of Family Growth**

interview for most items, but data were to be collected for a selected set of sensitive items using audio computer-assisted self-interview (ACASI) in which the respondent listened to a prerecorded audio reading of the questions via a headset connected to the laptop and entered responses directly into the laptop.

- Signed informed consent was to be required for every selected eligible respondent. Minors 15–17 years of age were to be required to have both the signed consent of a parent or guardian and their own signed assent. Interviews obtained without signed consent were not to be allowed. (The interviewer was to be permitted to sign for an adult respondent 18–44 years of age, but this was to be done infrequently.)
- Interviewers were to ask questions about fertility, contraceptive use, sources and types of family planning services, and maternal and child health using structured questions read by the interviewer from the laptop display or from audio recordings heard on a headset.
- The target interview completion rate for females was to be 80 percent of the expected number of selected eligible women and 75 percent of the expected number of selected eligible men, where the interview completion rate is the product of household screener response rates and the main interview response rate.
- The contractor, in cooperation with NCHS, was to design and implement procedures for measuring and controlling the quality of data collection and preparation procedures. These included verification of a sample of interviews.

# Sample Design

## Summary of Sample Design

This section summarizes the sample design and sample release features of NSFG Cycle 6. A more detailed account of the sample design follows this summary. The NSFG Cycle 6 sample design consisted of four stages of selection to choose eligible sample persons. Women, teenagers 15–19 years of age, and black and Hispanic persons were selected at higher rates, yielding an oversample of such persons in Cycle 6. The sample design is summarized in figures 2 and 3.

The entire land area of the United States was divided in the 1990 census into 3,141 counties and county

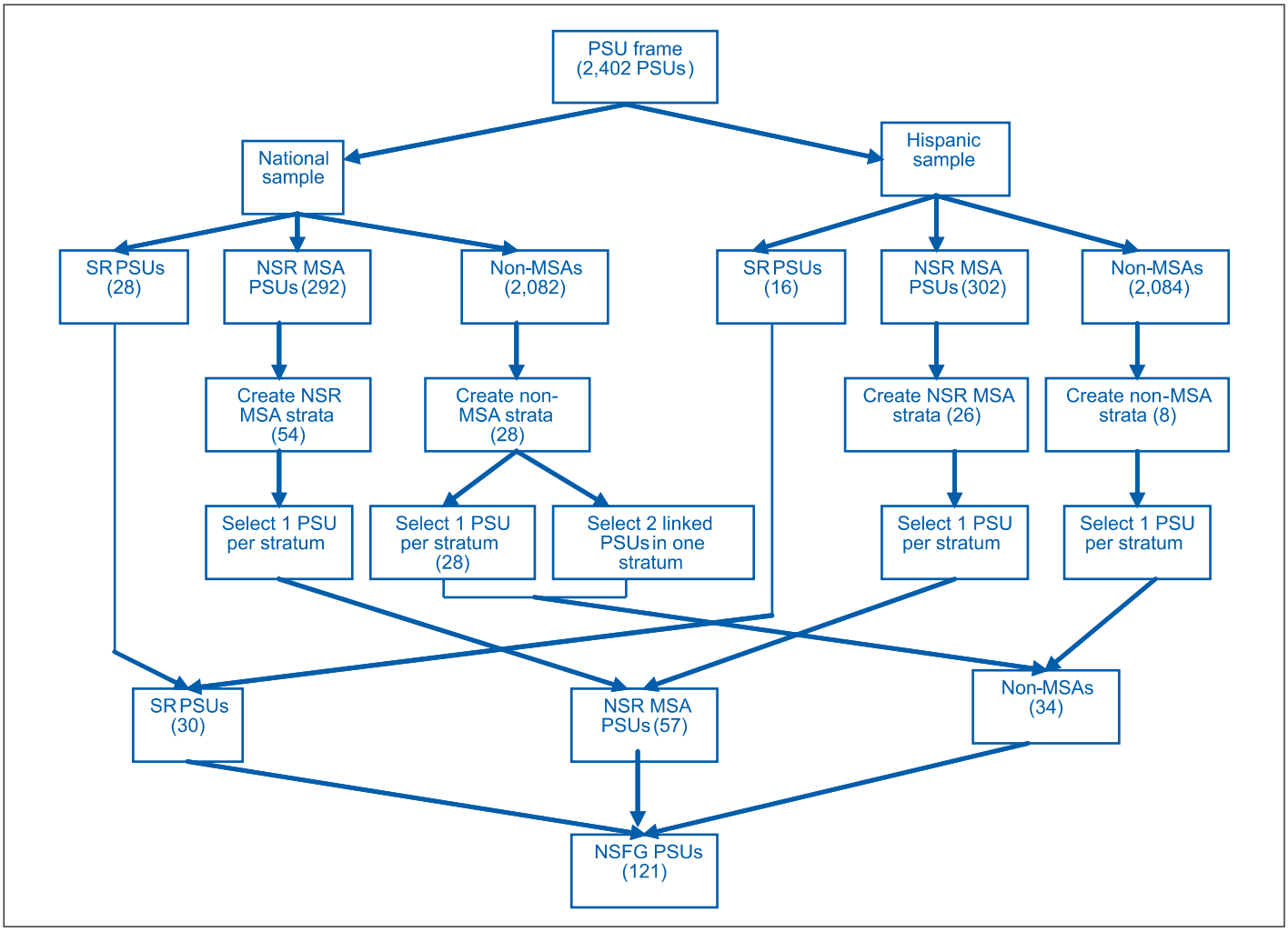


**Figure 2. National Survey of Family Growth Cycle 6 selection of primary sampling units and segments**

equivalent units. The metropolitan counties and other units were grouped to form 2,402 PSUs. These PSUs were partitioned into three major strata—large metropolitan areas, other metropolitan areas, and nonmetropolitan areas. The PSUs within each major stratum were further grouped by geography and population size.

Two separate samples of the 2,402 primary areas were selected for Cycle 6. A national sample selected 110 PSUs from the 50 States and the District of Columbia. The national sample PSU selection was based on 1990 census counts of occupied housing units (households) in each county or other unit. This national sample was efficient for obtaining a sample of black, white, and other race households. It was not efficient for selecting a sample of Hispanic households.

A second national sample of primary areas was selected to supplement the sample of Hispanic households in the United States obtained in the national sample. Since the selection of the Hispanic primary area sample was completed after the 2000 census individual-level redistricting data were released but before 2000 census household data were available, the counties were selected using counts of Hispanic persons rather than occupied housing units. The same 2,402 primary areas were grouped into the same three partitions used in the national sample, and each partition was stratified by geography and the number of Hispanic persons. Within each stratum, one primary area was selected as the PSU to represent the stratum. The selection procedure used a method developed by Kish and Scott (7) to maximize the probability of retaining a primary area already selected as a PSU in the national sample with probability proportional to the year 2000 Hispanic population of the PSUs. The selection to maximize overlap yielded a total of 39 Hispanic sample PSUs that were already in the national sample and 11 additional Hispanic PSUs.

Within the 110 + 11 = 121 national and Hispanic sample PSUs, a database of population counts for census blocks was obtained from the 2000 census. Information was available on block population by race and ethnicity.

In the second stage of selection, census blocks were stratified into four domains within each PSU:

1. Nonminority
2. More than 10 percent black persons but less than 10 percent Hispanic persons

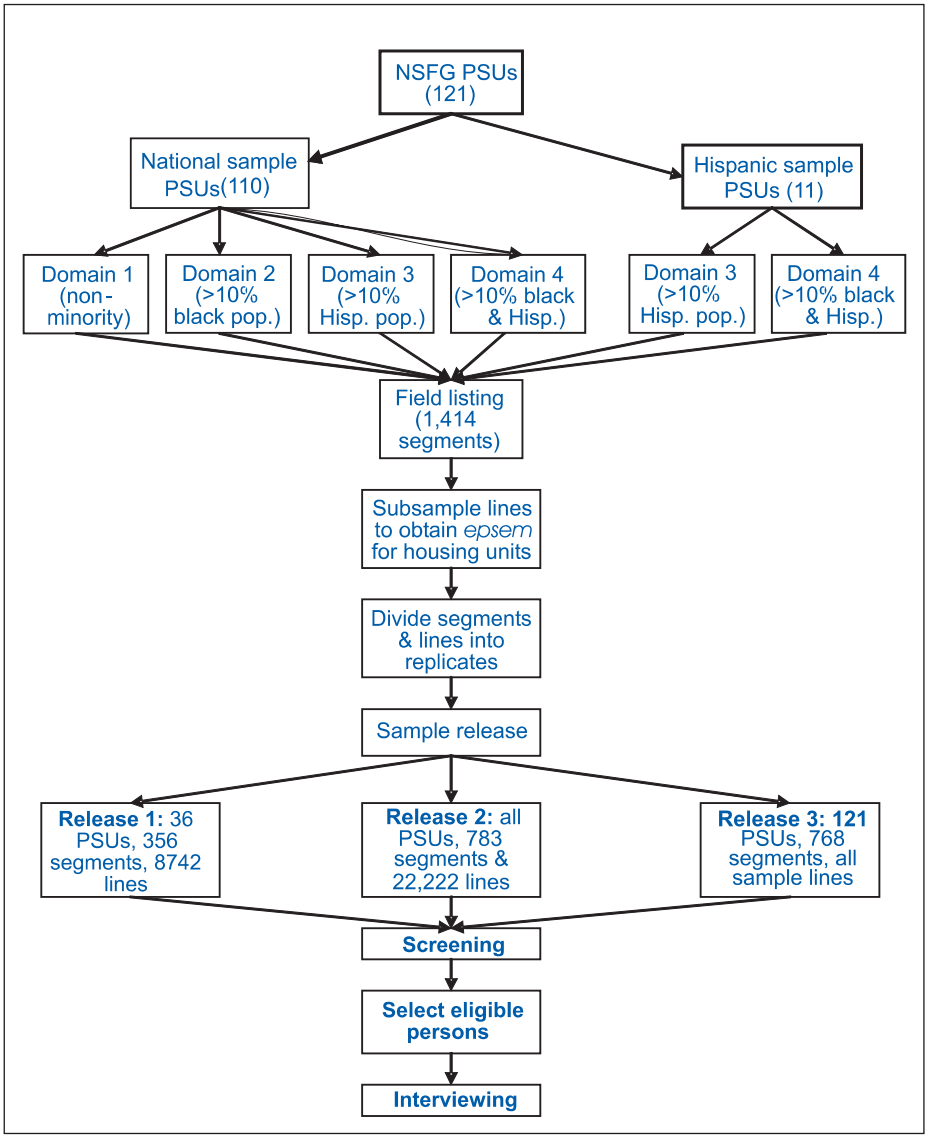


**Figure 3. Selection steps in sample implementation for Cycle 6 of the National Survey of Family Growth**

3. More than 10 percent Hispanic persons but less than 10 percent black persons
4. More than 10 percent black persons and more than 10 percent Hispanic persons.

Blocks were then selected within each domain with probabilities proportionate to the estimated number of 2000 census households.

Some blocks had only a few housing units and were not adequate to support subsequent required interviewer workloads. Such small blocks were linked together with other larger blocks, and sometimes with other smaller blocks, to form "segments" that were large enough to support data collection workloads. In urban areas, the minimum size for a block before linking was required was 75 estimated households while in rural areas the minimum was 50. The segments thus formed consisted of entire blocks or set of linked blocks. A total of 1,414 segments were selected across the 121 PSUs.

Trained household listers were sent to each of the sample segments selected in the second stage to list housing units on the blocks in the segments. Housing unit lists were returned to the ISR and keyed into a file containing one line per housing unit.

The third stage of selection chose sample lines (housing units) from each sample segment with probabilities designed to obtain equal chances of selection for households within segment domains. Sample lines (sampled households) in census blocks in which more than 10 percent of the population was black or Hispanic (domains 2, 3, and 4 shown previously) were selected at higher rates to increase the number of black and Hispanic persons in the sample.

The fourth stage was the selection of eligible persons within sample households. Interviewers visited housing units represented by each selected sample line or address. When the housing unit was occupied, interviewers attempted to list all persons living there. One eligible person was chosen randomly in every sample household that contained one or more eligible persons. The within-household selection used preassigned measures of size that varied by gender, age group, and race and ethnicity. Teenagers 15–19 years and females received larger measures of size to increase their chances of selection within the household and to yield larger sample sizes required for the target sample sizes for these groups. Persons within households were then selected with a probability proportionate to the measure of size.

Persons living away from home in college or university dormitories, sororities, and fraternities were considered to be part of their parents' household and were listed in them. If a college student was chosen, she or he was transferred to an interviewer working in the nearest PSU to the college or university for interviewing. In some cases, interviews were obtained at the sample household when selected students returned home for summer or holidays.

A two-phase sampling procedure was used to obtain the target number of interviews within the survey budget. Five weeks before the end of data collection, a total of 11,721 interviews had been completed. Not all sample lines that had been released were successfully screened to determine if eligible persons resided at the address. Additional sample lines had eligible persons who were selected for interview, but the interview had not been completed. Among these remaining lines to be screened and eligible selected

IntroSC
SC-1 I am (name) from the University of Michigan. Here is my identification card. We are conducting the National Survey of Family Growth for the National Center for Health Statistics, which is part of the US Public Health Service.

This letter explains all about this survey. You may remember receiving this letter and a brochure in the mail recently. Please take some time to read this important information.

Do you have any questions about anything (you have read/I have read to you) about the National Survey of Family Growth?

*Enter (1) to continue*
*Underlying range: 1 to 1, NODK, NORF*

| | Name | Sex | Age | Race | Hisp | DormRes |
|---|---|---|---|---|---|---|
| HL(1) | | | | | | |
| HHL(2) | | | | | | |
| HHL(3) | | | | | | |
| HHL(4) | | | | | | |
| HHL(5) | | | | | | |
| HHL(6) | | | | | | |
| HHL(7) | | | | | | |
| HHL(8) | | | | | | |
| HHL(9) | | | | | | |
| HHL(10) | | | | | | |
| HHL(11) | | | | | | |
| HHL(12) | | | | | | |
| HHL(13) | | | | | | |
| HHL(14) | | | | | | |
| HHL(15) | | | | | | |

**Figure 4. Within-household screener and household roster**

persons remaining to be interviewed, a second-phase sample was chosen. The second-phase sample consisted of approximately 1,700 lines that had not been screened and another 1,500 households containing eligible persons but without main interviews. This size was chosen to yield a target of 12,500 completed interviews without exceeding the funds that remained to pay for interviewer labor for data collection. The second-phase sample yielded an additional 850 completed interviews. The final response rate needs to be weighted to account for the second-phase sampling rates. The final weighted response rate for men and women combined was 79 percent, virtually the same for women (80 percent) and men (78 percent). The detailed description of the design follows.

## Sample Size Requirements

The final sample size for Cycle 6 of the NSFG was determined on the basis of the resources and interviewer hours available at the time of sample selection. The target sample size was set at 12,500 completed interviews. The actual sample size was 12,571 completed interviews.

Among the 12,500 target interviews, at least 2,000 completed interviews were required with teenagers 15–19 years of

age. At least 2,000 completed interviews were required with Hispanic eligible persons and at least 2,000 with black eligible persons. In addition, at least 7,500 interviews were required with female respondents.

These requirements represent substantial oversamples of teen, black, and Hispanic populations, and a modest oversampling of females. To achieve these oversample targets, three sample design features were employed.

First, a national sample of Hispanic households was added to a national sample of all households to increase the number of completed interviews with Hispanic eligible persons in the final sample. Second, within sample PSUs, second-stage sampling rates were set to increase the number of black and Hispanic households in the sample. Four strata of blocks were created in sample PSUs, and different sampling rates were set across each of the strata. Lastly, within households, teens and females were assigned higher probabilities of selection. This was achieved through the assignment of measures of size to each eligible person in the household. A probability proportional to a measure of size was used to select one eligible person in each household (see figures 4 and 5 for an illustration of the within-household selection process).

The measures of size were larger for individuals in target groups within the household, such as teens 15–19 years of age or females and lower for other groups such as individuals 25–44 years of age. For example, in the black household shown in the second table in figure 5, there are six persons in the household. Two are not eligible because they are too young. They received a measure of size equal to zero, which means that they cannot be selected for the study. Among the remaining four persons, the 15-year-old female receives a measure of size from the first table in figure 5 of 1.00. The 21-year-old male receives the next highest measure of size, 0.90. The remaining two eligible persons, a 40-year-old female and 42-year-old male, receive correspondingly smaller measures of size.

In the selection, the measures of size are then cumulated, and a random number is generated by the questionnaire software between zero and the total sum of the measures of size (in this case, 3.45). The random number is then compared with the cumulative sum of measures of size to determine which interval it falls into. For example, in the illustration the random number 0.95 falls in the interval for the first eligible person, the 15-year-old female. The 15-year-old female is selected for the sample.

The measures of size were determined before data collection began and were the same for every household in the survey. There were 18 possible measures of size (see figure 5). The original sample design specified the number of completed interviews in each of 18 subgroups defined by age interval (15–19, 20–24, and 25–44), gender (female and male), and race and ethnicity groups (black, Hispanic, white or other). The measures of size were selected to yield a prespecified number of teens in the sample through a series of simulated sample selections applied to data from the Current Population Survey (CPS). The remaining sections in this chapter describe the selection procedures in greater detail.

| Age and gender | Black | Hispanic | White/ other |
|---|---|---|---|
| Female: | | | |
| 15-19 | 1.00 | 1.00 | 0.84 |
| 20-24 | 0.94 | 0.94 | 0.84 |
| 25-44 | 0.87 | 0.84 | 0.28 |
| Male: | | | |
| 15-19 | 1.00 | 1.00 | 0.84 |
| 20-24 | 0.90 | 0.57 | 0.47 |
| 25-44 | 0.68 | 0.47 | 0.12 |

Illustration:

| Race and ethnicity | Gender | Age | Measure of size | Cumulative measure of size | Random number (from 0 to 3.45) |
|---|---|---|---|---|---|
| Black | Female | 6 | 0.00 | --- | |
| Black | Female | 15 | 1.00 | 1.00 | ← 0.95 |
| Black | Female | 40 | 0.87 | 1.87 | |
| Black | Male | 10 | 0.00 | --- | |
| Black | Male | 21 | 0.90 | 2.77 | |
| Black | Male | 42 | 0.68 | 3.45 | |

---Category not applicable.

**Figure 5. Within-household measures of size and illustration of within-household selection**

## First-Stage Selection of Primary Sampling Units

### Sampling frame

The Census Bureau in 1990 divided the land area of the United States into 3,141 counties and county equivalents (Louisiana parishes; Alaska boroughs and census areas; independent cities in

Maryland, Virginia, Missouri, and Nevada; and the District of Columbia). Metropolitan areas were sampled as one PSU, even if they included more than one adjacent county. The counties used for each metropolitan area were defined by the U.S. Office of Management and Budget in 1990 as ''primary metropolitan statistical areas'' (MSAs). Each nonmetropolitan county was treated as a potential PSU.

The ISR grouped these metropolitan and nonmetropolitan units into 2,402 PSUs in its national area probability sample. The ISR used these PSUs as the sampling frame for Cycle 6 of the National Survey of Family Growth.

Data from the 2000 census were just becoming available at the time of NSFG Cycle 6 selection. However, complete information needed for PSU identification and selection were not available at the time the sample had to be chosen. The frame used units and population data from the 1990 census of population and housing for selection of PSUs in the national NSFG Cycle 6 sample.

The national sample would not yield a large enough sample of Hispanic households to meet sample size goals without substantial oversampling of second-stage sampling units. A second national Hispanic PSU sample was chosen to supplement the national sample. It used the same frame of 2,402 PSUs in the national sample, but the stratification was altered to increase the number of Hispanic households available in sample PSUs.

The national sample chose a sample of 110 PSUs using 1990 census data. PSUs were selected with probabilities proportional to all 1990 households, Hispanic and non-Hispanic. The Hispanic sample selected 50 PSUs from the same frame using 2000 census data on Hispanic persons. Hispanic sample PSUs were selected with probabilities proportional to the number of Hispanic persons counted in the 2000 census. The Hispanic PSU sample also used a technique to maximize overlap with the national sample of PSUs. The Hispanic sample chose 39 PSUs that were already in the national sample and another 11 PSUs that were not in the national sample. These 11 additional PSUs were added to the national sample to yield a final sample size of 121 PSUs. Both PSU samples are described in the next two sections.

## National Sample

The national sample component was drawn from the ISR's national sample of primary sampling units. Metropolitan areas vary substantially in size, from as few as 25,000 to as many as 3.2 million households in the 1990 census. The 28 largest MSAs each had more than 600,000 households. These MSAs were so large that they were chosen with certainty for the sample. These 28 MSAs constitute 28 strata in the NSFG national sample with exactly one PSU in each. Since they are chosen with probability one, they are referred to as self-representing (SR) units.

The remaining 2,374 non-self-representing (NSR) PSUs were divided into 290 metropolitan and 2,084 nonmetropolitan PSUs. Each of these partitions was then grouped into a combined set of 82 strata that contained approximately the same number of households (about 721,000 households in a NSR stratum).

This stratification of PSUs was based on several criteria. PSUs were first grouped into one of nine census geographic divisions defined by groupings of States. In addition, the States of Alaska and Hawaii were separated from the Pacific division and treated as separate strata of PSUs. Within divisions, PSUs were grouped into a total of 52 MSA and 30 non-MSA strata. In Alaska and Hawaii, the distinction between MSA and non-MSA PSU was not made, and each State was treated as a stratum.

Within a division and MSA category, PSUs were further stratified on the basis of size and geography. For example, in the Middle Atlantic Division, MSA PSUs were grouped into eight strata, ranging from large MSAs in one State to small MSAs in the western part of two States.

A single PSU was then selected from within each of the 82 strata. For purposes of minimizing the need to hire new interviewers, a procedure to maximize overlap between two samples (7) was used to determine which PSUs from the 1980 ISR national PSU sample would be retained in the 1990 sample. The selection was random and with probabilities proportional to the 1990 households in the PSU. A diagram summarizing the national and Hispanic sample selection is presented in figure 2.

## Hispanic Sample

The Hispanic sample stratification followed a procedure similar to that used in the national sample. The national sample's 2,402 PSUs were the starting point for the selection. Each PSU was evaluated on the basis of its 2000 census Hispanic total population. A minimum of 280,000 Hispanic persons was established as the required size for SR PSUs. Fourteen MSAs met this criterion. Two adjacent PSUs met the minimum size threshold when they were combined. For the sake of sample efficiency, these two PSUs were treated as SR PSUs, resulting in a total of 16 SR PSUs.

The remaining 302 MSAs were treated as NSR PSUs. These were grouped into 26 strata primarily by combining 1990 national sampling frame strata within census divisions and by geographic proximity. The 26 strata were created to make the numbers of Hispanic persons across them as equal as possible, with populations ranging from 378,000 to 533,000.

The 2,084 non-MSA PSUs were grouped into eight strata, initially within census divisions and retaining geographic proximity. Because of the concentration of Hispanic persons in the south and west regions, however, there is one non-MSA stratum covering the entire northeast and midwest regions, and one covering the south Atlantic and east south central divisions. Three strata were created in the west south central division and three in the west region. (In this case, the mixed Alaska and Hawaii strata from the national sample were combined with the Pacific coastal non-MSA strata and treated as a non-MSA.) Hispanic populations in these non-MSA strata ranged from 235,000 to 462,000. Among the 16 SR PSUs, 14 were already selected in the

NSFG national sample. Thus, the Hispanic sample added two new SR PSUs for the NSFG Cycle 6 sample.

In the 26 NSR non-MSA strata, a Kish-Scott procedure (7) was used to select PSUs for the Hispanic sample. The goal was to maximize the overlap between the national and Hispanic NSR non-MSA PSUs. In the procedure, units in one sample (the national sample) are first selected to determine which will be 'retained' in the second sample. In the strata where multiple PSUs could be retained from the national to the Hispanic sample, one PSU was selected from those eligible for retention with a probability proportional to the 2000 census Hispanic population. Where only one could be retained, it was selected with certainty. Where none could be retained, a new PSU was selected from among all PSUs in the stratum with probabilities proportional to the Hispanic population. New PSU selections not in the NSFG national sample were made in 5 of the 26 NSR non-MSA strata.

In the eight non-MSA and mixed Alaska-Hawaii strata, the same Kish-Scott procedure was used to select PSUs. Four national sample PSUs were retained and four new NSR non-MSA PSUs were selected. In summary, a total of 11 new PSUs were added to the national sample from the Hispanic sample. With these new PSUs, the combined NSFG national and Hispanic samples contained 30 SR, 57 NSR MSA, and 34 NSR non-MSA PSUs, or a total of 121 PSUs.

## Second-Stage Selection of Segments

### Stratification of Blocks

As noted previously, the NSFG Cycle 6 design requirements specified substantial oversampling of black and Hispanic men and women across the 15–44 years of age range. In order to achieve the target oversampling, probabilities of selection of second- and third-stage units—segments and households, respectively—were varied across domains created within each PSU.

The second-stage sampling units were census blocks, or combinations of census blocks that had sufficient numbers of households to sustain efficient survey data collection. For each of the 121 NSFG Cycle 6 PSUs, a list of blocks was obtained from the 2000 census data.

Census blocks have a varying number of households and eligible persons. Some blocks have few, if any, households. These small blocks pose problems for efficient screening and interviewing in any study. The sample design specified that in order to obtain units of minimum sufficient size for efficient data collection, blocks should be linked to one another into units called segments with at least 75 households in urban areas or 50 households in rural areas. Linking was within block group boundaries to increase the chance that geographically contiguous blocks were linked to form segments. These segments were the second-stage sampling units.

At the time of selection, only limited information on population and housing from the 2000 census was available for blocks for all PSUs. The 2000 census blocks were used as second-stage sampling units. Numbers of black, Hispanic, and white or other race households were not available at the time of sample selection. The number was estimated by dividing the total black, Hispanic, and white or other population for the segment by the national average number of persons per household (obtained from the March 1999 CPS) for the respective group. This division yielded an estimated number of households for each group in the group in the block.

Blocks and segments within PSUs were grouped into one of four domains within each PSU based on the percentage of black or Hispanic households in the segment. Domain 1, the nonminority stratum, consisted of those segments with less than 10 percent black or 10 percent Hispanic persons. Domain 2 contained segments with at least 10 percent black persons, but less than 10 percent Hispanic persons. Domain 3 contained segments with at least 10 percent Hispanic persons but less than 10 percent black persons. Domain 4 contained segments with at least 10 percent black and at least 10 percent Hispanic persons. All race and ethnicity prevalence rates were computed at the census block group level to reduce classification of geographically contiguous census blocks into different domains.

Sampling rates for households within each domain were determined to achieve target numbers of completed interviews with black and Hispanic households. During the sample planning, the final sample size could not be precisely fixed. Limitations in funding forced consideration of several different target sample sizes, and target sizes for various age, gender, and race and ethnicity groups. For purposes of second- and third-stage selection, sampling rates within PSUs were established before final sample size targets were known.

Initial domain sampling rates were established to yield sufficient sample sizes of minority groups under several different NSFG sample size targets. Higher rates were needed for Domains 2, 3, and 4 in order to achieve target subgroup sizes under all final sample sizes considered. The higher sampling rates in these three domains meant that there would be a consequent increase in sampling variance and loss of effective sample size through the weighting needed to compensate for unequal sampling rates.

After simulation of many design alternatives, and examination of sample sizes in target groups and effective sample sizes, sampling rates for Domains 2, 3, and 4 were set relative to Domain 1. The sampling rate for Domain 2 was 80 percent higher than the sampling rate for Domain 1, while rates for Domains 3 and 4 were 110 percent higher (more than double) the rate for Domain 1. These rates are summarized in table A.

The number of segments to be selected from each PSU and domain depended on both the expected increase in variance due to cluster sample selection within the PSU and the cost of traveling to segments within a PSU. An analysis of the variance and cost trade-offs led to the conclusion that a total of 1,400 segments were to be

**Table A. Allocation of sample segments and sampling rates within domains: National Survey of Family Growth, Cycle 6**

| Domain | Number of segments allocated | Number of segments listed | Sampling fraction |
|---|---|---|---|
| 1. Nonminority | 618 | 614 | 1 in 2,572 |
| 2. More than 10 percent black | 269 | 264 | 1 in 1,429 |
| 3. More than 10 percent Hispanic | 288 | 287 | 1 in 1,225 |
| 4. More than 10 percent black and Hispanic | 239 | 219 | 1 in 1,225 |
| Total | 1,414 | 1,384 | 1 in 1,788 |

selected across all 121 PSUs.

The number of segments within a PSU and within a domain was allocated based on the number of households in the domain in the PSU. However, no households were to be selected from Domain 1 and 2 in the 11 additional Hispanic sample PSUs. That is, sample segments were to be drawn from Domains 3 and 4 only in the 11 additional Hispanic sample PSUs. Thus, the Hispanic sample would increase the number of Hispanic households and eligible persons in the sample.

The allocation of segments across domains had to be rounded to a whole number in each domain and PSU. While rounding could have been controlled across the domain-PSU cells of the allocation, it would have been a time-consuming task with little real benefit in terms of the final sample size. The rounding was not controlled and a total of 1,414 segments were allocated across the PSUs, slightly more than the target number.

## Selection of Segments

Once minimum measures of size for each segment and the number of segments per domain in a PSU were determined, a sample of segments was selected from each PSU. Within a domain and PSU, the census blocks were ordered geographically by tract, block group, and block number.

Segments were selected within a PSU using a systematic method with probabilities proportional to the 2000 census count of estimated households in the segment. The cumulative number of households for each segment in the domain was calculated, cumulating in list order. The sum of domain household counts in the PSU was divided by the number of segments allocated to the domain to obtain a sampling interval for selection. A random number was chosen for each domain in each PSU that was between zero and the sampling interval for the domain. The first selected segment was identified by finding the first segment with cumulative size that exceeded the random number. The second selection was obtained by adding the sampling interval to the random start and identifying the segment with cumulative size that first exceeded the sum. The selection continued, adding the interval to the last sum to find the next selection until the domain allocation was achieved.

The combination of the geographic order and systematic selection produced an implicit geographic stratification of segments within each PSU and domain. The implicit stratification was expected to yield a decrease in sampling variance for estimates obtained from the final sample results.

## Field Listing of Segments

Following segment selection, a folder was prepared for each segment containing maps and blank listing forms. The maps were generated through commercial software using the Census Bureau's TIGER (**T**opologically **I**ntegrated **G**eographic **E**ncoding and **R**eferencing system) files to delineate block boundaries. Instructions were prepared for listing each of the blocks in a segment, including a starting place to begin the listing and a direction to proceed around the block.

Segment folders were sent to experienced field interviewers for listing. Interviewers entered onto a segment housing unit listing sheet the number and street name of all housing units on each block. When no address was available, a description of the housing unit was entered onto the listing form. These lists contained the units in the third-stage selection of housing units.

The listing forms were returned to the ISR in Ann Arbor, Michigan, where they were checked for completion and then keyed in listing order to a computer file of listings. More than 140,000 addresses were listed in the sample segments.

Because the estimated numbers of households were used as a measure of size in the selection of segments, there was a large discrepancy between estimated size and actual count for 30 of the 1,414 segments. Several of these segments contained institutions such as prisons, jails, or college dormitories that had populations but no households, segments that could not be identified in advance from the available 2000 census materials. These 30 segments were treated as foreign elements in the segment list, and were excluded from subsequent within-segment listing and housing unit selection. Thus, the total number of segments in the 121 PSUs was 1,384.

# Third-Stage Selection of Housing Units

The third-stage selection of housing units was made from the keyed list of housing units. The individual listings were referred to as sample lines. In order to identify which sample lines would be selected and assigned for field data collection, a selection was required that would achieve the specified within-domain sampling rate for lines, housing units, and households. A probability of selection of the PSU and the segment was recorded for the first two stages of sampling. Given the target sampling rates shown in table A, a probability of selection for lines within each of the 1,384 segments was calculated. The lines were then subsampled at the specified rate systematically from the ordered list of lines in the segment. The combination of systematic selection and list order implicitly stratified the selection of lines

to provide representation of lines across blocks and the geographic area represented by the segment. A total of 74,132 lines were subsampled from the 1,384 segments.

Not every line selected for the sample was released for subsequent screening and within-household selection of eligible persons. In order to control sample size and constrain costs at or below budget, the final sample was released in two phases. In the first phase, operational constraints allowed a reduced interviewer staff to be trained. A sample of segments from a one-quarter sample of 33 NSFG PSUs was released. In the second phase, when the complete interviewer staff was trained across all PSUs, sample segments were released across all 121 PSUs. Each phase used a sample drawn from the available NSFG PSUs, segments, and lines.

A third phase of sample selection was employed at the end of the data-collection period. The phased sampling is described in greater detail in a subsequent section.

## Screening and Missed Housing Units

The selection of households was carried out through household screening in the field. Screening consisted of a short questionnaire administered at the doorstep of every occupied sample line released for the sample.

Based on ISR experience with listed addresses, 15 percent of sample lines were expected to be unoccupied or not actually housing units (for example, a housing unit converted for a commercial purpose). Based on estimates obtained from the March 1999 CPS, an additional 38 percent of households was expected to have no eligible persons. Screening was the process used to determine whether the sample line contained a household and whether any eligible persons resided in the household.

During the screening phase of the survey, interviewers were trained to check for housing units that were not found during listing. Missed housing units may occur when a listing interviewer overlooked a structure with one or more housing units or a part of a structure that was a separate housing unit. Missing units may also have occurred if a housing unit was constructed since the listing took place.

Interviewers were equipped with a laptop computer and a sample management system that contained all sample lines in the segment and highlighted sample lines they were to attempt to screen within the segment. The procedure for handling missed housing units in the field was as follows:

- At each sample line, interviewers checked to be sure that all housing units in the structure were on the sample list. Interviewers were to check mail boxes, doors, or utility meters that might indicate a unit present that was not listed. They were also instructed to ask screener respondents about any additional housing units in the structure where they were living.
- If an additional housing unit was discovered, and it was between the sample housing unit and the next listed unit, the interviewer was to add the additional unit to the online system and attempt a screening interview with the additional unit.
- When more than two housing units were missed between the sample line and the next listed housing unit, interviewers were instructed to suspend work, including contact with the household, and to call the ISR's sampling unit to receive further instructions. Before calling the sampling unit, the interviewers were to obtain a list of all additional housing units associated with the sample line.

If a large number of additional housing units were found, a subsample of the original sample line and the additional housing units was selected. This subsampling disrupts the equal probability of selection of households within each domain. Weights to account for the missed housing unit subsampling were incorporated into compensatory weights to adjust for unequal probabilities of selection.

# Fourth-Stage Selection of Sample Persons

The last stage of sample selection was conducted within the household during the screening activities. An adult member of the household was asked to provide a list of all persons living in the household. Information on the gender, age, and race and ethnicity of each person was recorded in the screening portion of the interview (see figure 4). Interviewers asked additional questions to be sure no one was missed, particularly college students living away from home at a dormitory, fraternity, or sorority. (College students living away from home in their own apartment or housing unit were covered by the household frame and were not considered to be part of their parents' household.) Dormitory residents were indicated on the household listing.

If no one in the household was between the ages of 15 and 44 years, the screening interview was terminated. If one or more eligible persons were found, the computer-assisted screening system made a selection of one eligible person in the household. That is, an eligible person was selected within each household.

## Subsampling by Age, Gender, and Race and Ethnicity

The sampling rates used within households with eligible persons varied by age, gender, and race and ethnicity. Since the selection was carried out by a computer application, a selection procedure that allowed person selection rates to be varied across sample releases could be implemented. This flexibility allowed SRC and NCHS staff to achieve target sample sizes more precisely in the face of uncertainty about rates of eligible persons in the population.

The original sample design specified the number of completed interviews in each of 18 subgroups defined by age interval (15–19, 20–24, and 25–44), gender (female and male), and race and ethnicity groups (black, Hispanic, white, or other). Figure 4 illustrates the household listing procedure and the upper panel of

figure 5 illustrates the configuration of the 18 cells.

To achieve the target sample sizes across these cells, varying rates of selection had to be set for each cell to account for expected eligibility rates for each cell. Table B presents expected and achieved eligibility rates for each of the 18 target subgroup cells. Expected eligibility rates were computed as the product of household eligibility rates and within-household rates by age and race and ethnicity as determined from a concatenated March 1998, 1999, and 2000 CPS data set.

As shown in the "Total" column of table B, approximately 62 percent of the households were expected to have one or more eligible persons. Expected eligibility for some cells—for example, black males 15–19 and 20–24—was quite low, less than 1 percent. Given these low rates, study staff were concerned that slight departures in observed rates from the expected rates could mean that target sample sizes might not be achieved. An additional concern was that response rates across these 18 cells might vary.

A within-household selection procedure was devised that, when implemented in the computer-assisted screener application, could be modified during data collection to refine within-household selection rates to adjust for observed eligibility and varying response rates across cells. The selection procedure involved assigning a "measure of size" to each of the 18 cells. Each person in the household would then be assigned a measure of size based on their age group, gender, and race and ethnicity. The measures of size were developed through a simulated sample selection from the concatenated CPS data used to generate the expected eligibility rates. For each member of a household in the concatenated CPS data file, a measure of size was assigned, and a subject was selected. The average probabilities of selection for persons in each cell were then computed and used to determine expected sample sizes by cell. The measures were adjusted cell by cell in order to achieve target sample sizes by gender, age, and race and ethnicity group. Figure 5 presents the size measures used in the initial sample release for each cell.

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

The lower panel of figure 5 illustrates how the selection would be executed in a hypothetical household with six persons: the black female age 15 has a measure of size of 1.0, which exceeds the random number of 0.95, so she is chosen to be interviewed. By assigning larger measures to teens (that is, those ages 15–19 years) and females within the household, larger samples of these individuals could be achieved. The selection probabilities within the household depend on the distribution of measures of size for other eligible persons in the household as well.

## Sample Release

The NSFG Cycle 6 sample was not released at one time. Due to concerns about the project budget and about interviewer attrition over the course of data collection, interviewers were recruited and trained in three groups during the data-collection period. The first cohort of interviewers was trained in March 2002, the second in June 2002, and the last in September 2002. As each of the first two groups of interviewers completed training, they were assigned samples from two different releases. Each release consisted of replicate samples drawn from the overall sample of PSUs, segments, and lines that were available for the NSFG Cycle 6 sample. Each release was a replicate of the overall sample design, selected with known probabilities of selection from the available listings. It is thus possible to construct overall probabilities of selection for each sample line in the NSFG sample (see Weighting Procedures).

In addition, in the last month of study data collection, there were sample lines that had not been assigned a final result. These lines had either not been successfully screened (that is, it was not known whether any eligible persons lived in the housing unit) or had been screened, but no interview had been completed with a selected respondent. During the last month of data collection, these lines were subsampled to allow for a more intensive data collection effort for a sample of lines. This subsampling

**Table B. Expected and observed eligibility rates by sex, age, and race and ethnicity: National Survey of Family Growth, Cycle 6**

| Sex and age | Black | | Hispanic | | White or other | | Total | |
|---|---|---|---|---|---|---|---|---|
| | Expected | Observed | Expected | Observed | Expected | Observed | Expected | Observed |
| Female: | | | | | | | | |
| 15–19 years | 0.007 | 0.010 | 0.010 | 0.010 | 0.036 | 0.032 | 0.053 | 0.052 |
| 20– 24 years | 0.009 | 0.010 | 0.011 | 0.013 | 0.038 | 0.040 | 0.058 | 0.063 |
| 25–44 years | 0.039 | 0.048 | 0.044 | 0.048 | 0.163 | 0.144 | 0.246 | 0.240 |
| Male: | | | | | | | | |
| 15–19 years | 0.007 | 0.009 | 0.011 | 0.010 | 0.037 | 0.032 | 0.055 | 0.051 |
| 20–24 years | 0.006 | 0.006 | 0.009 | 0.010 | 0.027 | 0.028 | 0.042 | 0.044 |
| 25–44 years | 0.025 | 0.027 | 0.034 | 0.032 | 0.106 | 0.083 | 0.165 | 0.142 |
| Total | 0.093 | 0.110 | 0.119 | 0.123 | 0.407 | 0.359 | 0.619 | 0.591 |

and intensive data collection activity is referred to as the third phase of sample release.

The NSFG Cycle 6 national sample of PSUs was divided into subsets that would allow replicate samples consisting of one-quarter, one-half, two-thirds, and all PSUs. In addition, each of the 1,384 segments was assigned to 1 of 12 replicates through a systematic procedure within each domain. Lastly, the selected lines in each segment were also assigned systematically to 1 of 24 replicates. These three replication features allowed more refined management of sample releases throughout the study in order to achieve target sample sizes and avoid field costs exceeding available budget.

In the phase-one release, interviewers were recruited from PSUs in one of the four quarter-samples available for the national sample. The first training session was limited in size because only a quarter-sample in a limited number of PSUs was being released at first. The first phase consisted of a sample of segments and lines drawn from available listings in the one-quarter sample of PSUs.

The PSU quarter-sample contained 33 PSUs: the 8 largest SR PSUs and 25 NSR PSUs. The Alaska and Hawaii PSUs were also included in the first-phase sample. None of the 11 additional Hispanic sample PSUs were included. The first-phase sample was thus drawn from 35 PSUs.

Within the eight SR PSUs, two-thirds of the segments from each domain and 17 of the 24 replicate line samples in each segment were chosen. In the remaining 27 one-quarter sample PSUs, all segments and 17 of 24 replicate line samples were selected. The first-phase sample contained 358 segments and 14,107 lines, adequate for the size staff trained in the first phase.

The second-phase sample was selected for release starting in June 2002, and included PSUs from the national and Hispanic samples. A second larger cohort of interviewers was trained in June and September. An additional 425 segments were released across all domains in the remaining 121–35=86 national and Hispanic sample PSUs. (No additional sample was released in the original quarter sample in Phase 2.) A total of 14,443 lines were released in June and 4,524 lines in September. Thus, across the February, June, and September training groups, a total of 33,074 lines were released across all four domains in all 121 PSUs and 783 segments.

# Sampling in Three Phases

As the data collection process matured during September through January, preparations began for a third-phase sample release. This third phase was, as described previously, a "double or two-phase sample design" first defined by Hansen and Hurwitz (8) as a tool to reduce the nonresponse bias in survey statistics.

The elements of a "two-phase design" (as Hansen and Hurwitz used the term) are:

1. The design and implementation of a survey design on a large sample (labeled the "first phase")
2. The selection of a probability sample of the nonrespondents to the first-phase sample
3. The use of a different participation protocol for the second phase

There are two impacts of a two-phase design. First, if the second-phase protocol is successful in measuring 100 percent of the *sampled* nonrespondents from the first phase, nonresponse bias is eliminated. In practice, no subsample of nonrespondents attains a 100 percent response rate and thus some nonresponse bias remains. Second, the cases sampled into the second phase that are successfully interviewed are assigned new selection weights (reflecting the fact that they must "represent" the nonselected nonrespondents). This additional weight component generally increases the variance of the estimates.

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

In the NSFG Cycle 6 sample design, a subsample of nonrespondents was chosen as Phase 3 of the design. Following the Phase 1 release, study staff developed response propensity models to predict the probability that a given case would yield a completed screening interview or a completed NSFG interview (9).

Based on the propensity models, the 783 sample segments of Phases 1 and 2 were stratified on two major dimensions: the number of cases in the segment that were not finalized and the total expected propensities for active cases in the segment based on the propensity models (9, tables K and L). The distribution of segments on each dimension was divided into quartiles, "high," "medium-high," "medium-low," and "low." Across the two dimensions, the segments were thus classified into 16 subgroups, as shown in table C.

It was believed that the cell in table C that would yield the largest number of completed interviews in Phase 3 was the upper-leftmost cell. This cell contained 111 segments with high total propensities and large numbers of active cases. When visiting these segments the interviewers could visit several houses on the same trip, reducing cost per unit.

The Phase 3 sample was a stratified sample of segments, with all nonrespondent cases in a selected segment included in the sample. This cluster sample design was chosen based on cost model estimates computed during Phases 1 and 2 that showed that a large portion of the total cost of completing a case arose from travel cost to the sample segments.

**Table C. Number of Phases 1 and 2 segments by number of active incomplete cases in the segment at the end of Phase 2, and the estimated total propensity of active cases to complete an interview on the next visit: National Survey of Family Growth, Cycle 6**

| Estimated total propensity for active incomplete cases (Quartiles) | Number of active incomplete cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 111 | 51 | 26 | 7 | 195 |
| Medium-high (next highest ¼) | 55 | 65 | 48 | 28 | 196 |
| Medium-low (next highest ¼) | 30 | 51 | 66 | 49 | 196 |
| Low (smallest propensities) | 10 | 34 | 40 | 112 | 196 |
| Total | 206 | 201 | 180 | 196 | 783 |

**Table D. Phase 3 segment sampling fractions, by number of active cases in the segment, and the estimated total propensity of active cases in the segment to complete the interview on the next visit: National Survey of Family Growth, Cycle 6**

| Estimated total propensity for active incomplete cases (Quartiles) | Number of active incomplete cases at end of Phases 1 and 2 | | | |
|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low |
| High (top ¼ of propensities) | 0.58 | 0.58 | 0.58 | 0.58 |
| Medium-high (next highest ¼) | 0.58 | 0.58 | 0.29 | 0.29 |
| Medium-low (next highest ¼) | 0.29 | 0.29 | 0.19 | 0.19 |
| Low (smallest propensities) | 0.29 | 0.19 | 0.19 | 0.15 |

Sampling fractions across the strata were chosen to increase the relative numbers of segments that would contain a higher yield of completed interviews. The fractions varied by a factor of about 4.0 (table D) from the largest to smallest. This unequal chance design emphasized the cost efficiency of the Phase 3 design to produce interviews, not minimization of standard errors of the resulting data set.

The variation in sampling fractions over the 16 strata could add to an increase in the variance of estimates. Simulations prior to the selection of the Phase 3 sample suggested an increase in variance of approximately 20 percent. The highest selection probabilities were assigned to those segments with the largest total sum of propensities across cases in the segment, or the largest numbers of active cases. The smallest selection probabilities for the Phase 3 sample were assigned to segments with few active cases that had low propensities of being interviewed.

Under the theoretical perspective (10) guiding the NSFG's design and fieldwork, study staff sought a Phase 3 interview recruitment protocol that was distinctive from that used in Phases 1 and 2. Such a distinction is necessary (but not a *priori* sufficient) to attract sample persons who did not find the Phases 1 and 2 protocol effective, and thus increase response rates and reduce bias in the sample data. With the approval of two Institutional Review Boards and the Office of Management and Budget, the Phase 3 recruitment protocol involved the following components:

1. Use of the most productive interviewers on staff
2. Increased use of proxy respondents (such as neighbors) for the screening interview
3. A prepaid $5 incentive (versus no incentive) for cases that had not yet completed the screening interview
4. A prepaid $40 incentive for the main interview (versus a $40 incentive provided after the informed consent was signed)
5. A promised additional $40 incentive for a completed main interview

Phase 3 was limited to a 1-month period, following 11 months in the first two phases. The results of Phases 1, 2, and 3 are described in the next section.

# Sample Characteristics

## Response Rate Estimation

The target response rates for the NSFG were 80 percent for females and 75 percent for males, traditional response rates that do not reflect any of the unequal probabilities of selection of the sample design. The Cycle 6 design focused on 18 subclasses that were of key analytic interest, defined by age, race and ethnicity, and gender. Response rates for each of these groups were monitored throughout the data-collection period.

There are two complications for computing response rate estimates in the NSFG Cycle 6 survey:

1. The sample of housing units required screening to determine whether anyone 15–44 years of age was a member of the household. About 60 percent of the households screened had one or more eligible persons. However, sample units that were not successfully screened have unknown eligibility statuses.
2. A subsample of remaining nonrespondents in Phases 1 and 2 was drawn to represent all nonrespondents.

The American Association for Public Opinion Research guidelines (11) for computation of response rates specifies a computational form for response rates in such cases. Let the following indicator variables represent the status of various outcomes for each case:

Phases 1 and 2 screener cases ($n_1$):

$s_{i1}$ = 1, if the case is an eligible household (containing one or more persons 15–44 years of age); 0, otherwise;
$e_{i1}$ = 1, for eligible households with no screener; 0, otherwise;
$u_{i1}$ = 1, for households of unknown eligibility; 0, otherwise.

Phase 3 screener cases ($n_2$):

$s_{i2}$ = 1, if the case is an eligible household in Phase 2; 0, otherwise;
$e_{i2}$ = 1, for eligible Phase 2 household with no screener; 0, otherwise
$u_{i2}$ = 1, for Phase 2 household of unknown eligibility; 0, otherwise.

Phases 1 and 2 main interview cases ($n_3$):

$i_{i1}$ = 1, for completed interviews; 0, otherwise;
$m_{i1}$ = 1, for nonresponse to main interview; 0, otherwise.

Phase 3 main interview cases ($n_4$):

$i_{i2}$ = 1, for completed interviews in Phase 2, 0, otherwise;
$m_{i2}$ = 1, for nonresponse to main interview in Phase 2; 0, otherwise.

The Phases 1–2 response rate not reflecting household and person selection weights is:

$$R_1 = \left( \frac{\sum_{i=1}^{n_1} (s_{i1}+s_{i2})}{\sum_{i=1}^{n_1} [(s_{i1}+s_{i2}) + (e_{i1}+e_{i2}) + l\,(u_{i1}+u_{i2})]} \right) \bullet \left( \frac{\sum_{i=1}^{n_3} i_{i1}}{\sum_{i=1}^{n_3} (i_{i1} + m_{i1})} \right)$$

The Phase 3 response rate not reflecting household and person weights is:

$$R_2 = \left( \frac{\sum_{i=1}^{n_2} s_{i2}}{\sum_{i=1}^{n_2} [s_{i2} + e_{i2} + l(u_{i2})]} \right) \bullet \left( \frac{\sum_{i=1}^{n_4} i_{i2}}{\sum_{i=1}^{n_4} (i_{i2} + m_{i2})} \right)$$

Here $l$ in the denominator of the response rate is an estimated proportion of screener nonresponse cases of unknown eligibility (that is, cases for which the screener was not completed) that *do* contain one or more eligibles. We have estimated $l$ in two different ways:

1. Using the proportion of completed screening interviews that yielded households containing one or more eligibles; this is approximately 58 percent.
2. Using a statistical model reflecting the amount of effort required to contact the case and the tendency to be reluctant to the screening interview to predict whether the household contains an eligible person; this is approximately 64 percent.

It is standard practice to present "unweighted" response rates that do not account for differential probabilities of selection arising from oversampling demographic groups. In the case of double sampling designs, though, a weighting to account for the differential subsampling of the sample non-respondents remaining after Phase 2 should be applied to obtain an estimated rate that reflects the full survey design. Thus, an alternative to a completely unweighted rate is a partially weighted rate—one that accounts for the subsampling of nonrespondents at the end of Phase 2 but does not also weight to compensate for the varying selection probabilities for demographic group oversampling within Phases 1 and 2.

In other words, to estimate the response rate across the three phases, a Phases 1 and 2 response rate must be combined with the Phase 3 response rate. However, the Phase 3 response rate is only applicable to the group of cases subsampled for Phase 3. The overall response rate incorporating the Phase 3 response rates (which were based only on a subsample of the remaining nonrespondents of Phases 1 and 2) is:

$$R_1 + (1-R_1)\, R_2$$

The partially weighted response rates computed in this fashion appear in table E.

This response rate estimator is consistent with prior practice in the NSFG (as in Cycle 4, for example), but applied to the multiphase design of Cycle 6. The estimator reflects the sampling of a portion of the segments in the Phase 3 subsample, but does not reflect the unequal probabilities of selection applied to housing units and persons used to oversample black, Hispanic, and teenage respondents.

Another weighted response rate estimate can, of course, be constructed to address the oversampling of these demographic groups. Such a fully weighted response rate would estimate the response that might have been obtained if the entire target population were, under the same protocol, selected with no differential sampling at all. Groups oversampled in Cycle 6 include units in high density Hispanic and African-American neighborhoods, teenagers, and persons who live with smaller numbers of other age-eligible persons. A response rate estimate can be constructed that reflects all the selection weights associated with these factors, in addition to the Phase 3 probability of selection. This latter response rate is relevant to statistical analysis in which the selection weights are being used. It estimates the percentage of eligible persons who would respond if equal probabilities of selection had been employed in the sample selection.

There are two weights needed for computation of the fully weighted response rate. The first is the selection weight for the screener response rate, which includes only the household selection probabilities. All cases have a weight that reflects the probability of selection of the household. For Phases 1

and 2 cases,

$$w_{s1i} = \frac{1}{\Pr(\text{Household selection})}.$$

And for Phase 3 cases,

$$w_{s2i} = \frac{1}{\Pr(\text{Household selection})} \bullet \frac{1}{\Pr(\text{Segment selection in Phase 3})}.$$

For the main response rate, the selection weights also incorporate within-household selection probabilities. The Phases 1 and 2 selection weight for the main interview is the final product of all constituent selection weights (see the next section for detailed description of the weights):

$$w_{m1i} = \frac{1}{\Pr(\text{Household selected})} \bullet \frac{1}{\Pr(\text{Person selected} \mid \text{Household selected})}$$

The Phase 3 weight is:

$$w_{m2i} = \frac{1}{\Pr(\text{Household selected})} \bullet \frac{1}{\Pr(\text{Segment selected in Phase 3})} \bullet \frac{1}{\Pr(\text{Person selected} \mid \text{Household selected})}$$

The final response rate can directly combine the multiple phases:

$$R_w = \left( \frac{\sum_{i=1}^{n_2} (w_{i1i}s_{i1} + w_{s2i}s_{i2})}{\sum_{i=1}^{n_2} \left[ \begin{array}{l} (w_{s1i}s_{i1} + w_{s2i}s_{i2}) + \\ (w_{s1i}e_{i1} + w_{s2i}e_{i2}) + \\ l(w_{s1i}u_{i1} + w_{s2i}u_{i2}) \end{array} \right]} \right) \bullet \left( \frac{\sum_{i=n_2+1}^{n_4} (w_{m1i}i_{i1} + w_{m2i}i_{i_2})}{\sum_{i=n_2+1}^{n_4} (w_{m1i}i_{i1} + w_{m2i}m_{i2})} \right)$$

These fully weighted response rates are shown in table F. The fully weighted rates are lower than the partially weighted response rates in table E. This decline is primarily due to:

1. Lower response rates in the undersampled age groups (for example, 65 percent for 25–44 year olds using 0.64 eligibility for screener noninterviews compared with 73 percent for 15–19 year olds)
2. Lower response rates in the undersampled other race and ethnicity group (for example, 66 percent for other using 0.64 eligibility for screener noninterview compared with 73 percent for black persons)
3. Lower response rates among Hispanic persons outside of the supplemental Hispanic PSUs (not shown in table F).

For comparisons to previous NSFG Cycles, the response rates that are unweighted except for the Phase 3 subsampling (table E) are most appropriate. For consideration of response rate differences that might be applied in a fully weighted analysis of the Cycle 6 data, response rates in table F may be preferred.

The overall response rate at the end of Phases 1 and 2 was approximately 64 percent, using the AAPOR definition with an estimated eligibility rate among the nonrespondent screener cases (but not reflecting unequal probabilities of selection across persons). The Phase 3 response rate was approximately 40 percent of the 36 percent of cases remaining, which added about 14.4 percent to the 64 percent response rate at the end of Phase 2. Adding the 14.4 to the 64 percent yielded a combined response rate of between 78 and 79 percent, using the approved AAPOR double sample computation, not reflecting unequal probabilities within phases (11). In that sense, the Phase 3 sample was very effective in

**Table E. Response rates, weighted only for subsampling in Phase 3, but not for other probabilities of selection, using two estimates (0.64 and 0.58) of the proportion of nonscreened households that had an eligible respondent: National Survey of Family Growth, Cycle 6**

| Age and race | Overall Cycle 6 response rates, using .64 eligibility for screener noninterview | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | | Total |
| | Black | Hispanic | Other | Total | Black | Hispanic | Other | Total | |
| All ages | 81 | 82 | 77 | 79 | 82 | 78 | 75 | 77 | 78 |
| 15–19 years | 87 | 84 | 74 | 78 | 83 | 85 | 78 | 80 | 79 |
| 20–24 years | 82 | 82 | 79 | 80 | 80 | 80 | 79 | 79 | 80 |
| 25–44 years | 80 | 81 | 77 | 79 | 82 | 76 | 73 | 75 | 77 |
| Black | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 82 |
| Hispanic | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 80 |
| Other | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 76 |

| Age and race | Overall Cycle 6 response rates, using .58 eligibility for screener noninterview | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | | Total |
| | Black | Hispanic | Other | Total | Black | Hispanic | Other | Total | |
| All ages | 82 | 83 | 78 | 80 | 83 | 79 | 76 | 78 | 79 |
| 15–19 years | 88 | 85 | 75 | 79 | 84 | 86 | 79 | 81 | 80 |
| 20–24 years | 82 | 83 | 80 | 81 | 81 | 81 | 80 | 80 | 81 |
| 25–44 years | 81 | 82 | 78 | 80 | 83 | 76 | 74 | 76 | 78 |
| Black | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 82 |
| Hispanic | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 81 |
| Other | . . . | . . . | . . . | . . . | . . . | . . . | . . . | . . . | 77 |

. . . Category not applicable.

**Table F. Fully weighted response rates, adjusted for all unequal probabilities of selection, using two estimates (0.64 and 0.58) of the proportion of nonscreened households that had an eligible respondent: National Survey of Family Growth, Cycle 6**

| Age and race | Overall Cycle 6 response rates, using .64 eligibility for screener noninterview | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | | |
| | Black | Hispanic | Other | Total female | Black | Hispanic | Other | Total male | Total |
| All ages | 74 | 70 | 68 | 69 | 72 | 66 | 64 | 65 | 67 |
| 15–19 years | 78 | 77 | 69 | 72 | 78 | 74 | 73 | 74 | 73 |
| 20–24 years | 78 | 72 | 71 | 72 | 78 | 62 | 70 | 69 | 71 |
| 25–44 years | 72 | 68 | 66 | 68 | 69 | 64 | 60 | 62 | 65 |
| Black | ... | ... | ... | ... | ... | ... | ... | ... | 73 |
| Hispanic | ... | ... | ... | ... | ... | ... | ... | ... | 68 |
| Other | ... | ... | ... | ... | ... | ... | ... | ... | 66 |

| Age and race | Overall Cycle 6 response rates, using .58 eligibility for screener noninterview | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Female | | | | Male | | | | |
| | Black | Hispanic | Other | Total female | Black | Hispanic | Other | Total male | Total |
| All ages | 75 | 71 | 68 | 70 | 66 | 65 | 66 | 73 | 68 |
| 15–19 years | 79 | 78 | 70 | 73 | 75 | 74 | 75 | 79 | 74 |
| 20–24 years | 79 | 73 | 72 | 73 | 63 | 70 | 70 | 79 | 71 |
| 25–44 years | 73 | 69 | 67 | 68 | 65 | 61 | 63 | 70 | 66 |
| Black | ... | ... | ... | ... | ... | ... | ... | ... | 74 |
| Hispanic | ... | ... | ... | ... | ... | ... | ... | ... | 69 |
| Other | ... | ... | ... | ... | ... | ... | ... | ... | 67 |

... Category not applicable.

increasing the overall response rate in only one month.

## Nonresponse Bias and Phase 3 Sampling

The next question is whether the Phase 3 sample and its increased response rate produced any reduction in nonresponse bias. Evidence regarding the nonresponse bias characteristics of the multiphase design is necessarily indirect, because the characteristics of the people who did not respond to the screener are not known. The NSFG attempted to control achieved interview counts on 18 different subclasses defined by age, gender, and race and ethnicity groups. Examining the Phases 1 and 2 and overall response rates of the 18 subpopulations helps to assess whether the response rates for subgroups were approximately equal, or whether it was necessary to use Phase 3 to attempt to make them more similar.

Figure 6 presents the response rates at the end of Phases 1 and 2 and the overall response rates for the 18 subclasses. The coefficient of variation of the response rates in Phases 1 and 2 is 7.6 percent of the mean response rate; the same measure of variation for the overall rates is 4.4 percent, a large decrease in the variation in response rates. This can be interpreted as an indirect indication that the Phase 3 follow-up reduced nonresponse error associated with age, gender, and race and ethnicity differences.

Figure 7 shows that the variation is quite systematic by age of the sample persons. At the end of Phases 1 and 2, teenagers had a 6 percentage points higher response rate than the 25 through 44-year age group, with those 20–24 years in between these two. At the end of Phase 3, the difference in response rates among the three groups was just 3 percentage points.

If nonresponse bias were related to survey effort and age, one would expect a relationship between the overall response rate and fertility experience variables that are a function of age. That is, as response rates increase and the sample becomes older, sample estimates of fertility experience variables (ever had intercourse, ever had a child, etc.) would be expected to change. Figure 8 presents a display used during data collection to monitor the status of fertility and other variables in the NSFG (9). It shows the relationship between the numbers of calls to sample addresses and the estimated proportion of women who reported ever having sexual intercourse, for each call. The results in figure 8 also include data from the Phase 3 subsample interviews.

The horizontal line in figure 8 is the estimated proportion of women who reported ever having had sexual intercourse from the Phase 3 subsample cases. The lower, slightly increasing line is the plot of the cumulative proportion as the full sample is cumulated from lower to higher numbers of calls. For example, using the left axis for this smooth cumulative line, the estimated proportion for cases completed at the second call is approximately 0.61, but for cases with 10 or fewer calls, the estimate has risen to nearly its final level, around 0.67.

The irregular line above the cumulative line is the estimated proportion for cases obtained in Phase 3 with exactly the indicated number of calls. Because the rates at each call in Phase 3 are much higher than those for the cumulative line across all phases below, another axis has been added to the right for this call-specific estimated

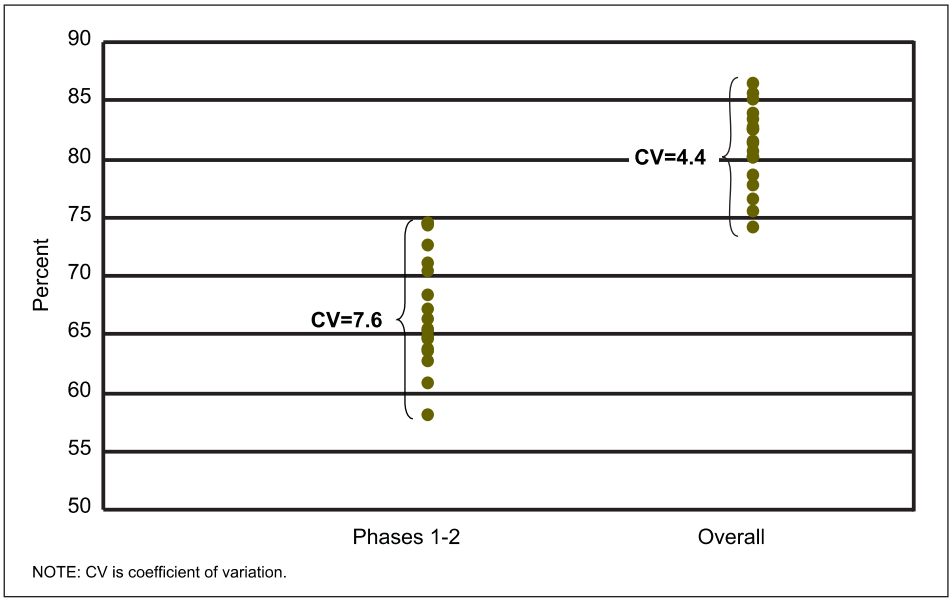


**Figure 6. Response rates for Phases 1 and 2 and overall for 18 subpopulations**

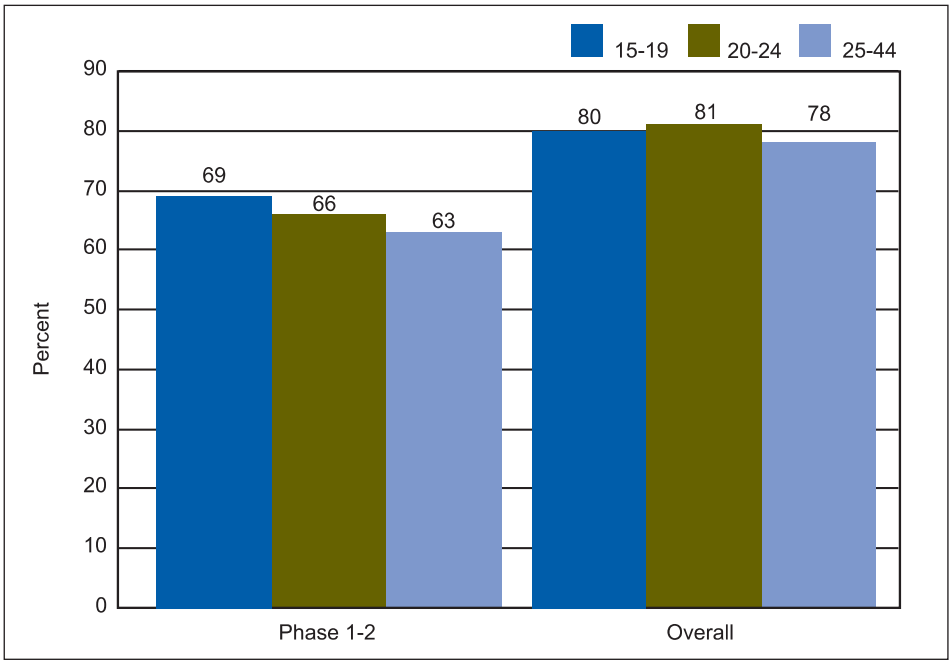


**Figure 7. Response rates for Phases 1 and 2 and overall (reflecting subsampling of nonrespondents) by age group; National Survey of Family Growth Cycle 6, 2002**

proportion for Phase 3. Those with interviews obtained on the first call had an estimated proportion near 0.95, while those with 20 calls had an estimated proportion near 0.70. Finally, near the right y-axis appears a single point that is the value of the final estimated rate, which includes the Phases 1, 2, and 3 data.

The proportion of females who ever had sex among Phase 3 respondents is 91 percent—higher than the estimate of about 88 percent for the first-phase respondents. This indicates that Phase 3 collected data for a different group of females who were more likely to have ever had sex than those interviewed in Phases 1 and 2.

These findings are consistent with findings of lower response rate variation by age after Phase 3. Phases 1 and 2 ended with a smaller than expected count of respondents older than 15–19 years of age. (It may be that the value of the $40 incentive used in Phases 1 and 2 is greater for a teenager than for an adult.) Phase 3 was more successful in interviewing older respondents. The effect on key statistics of this higher response rate in Phase 3 is that prevalence will be higher for those variables associated with older age and therefore with longer sexually active lives. That is, as the number of adults in the sample increases, due to higher response rates for older persons in Phase 3, the prevalence of sexual behaviors in the sample that cumulate over a lifetime will increase. From these simple analyses, there is evidence of reduced nonresponse error associated with the demographic characteristics because of the Phase 3 design. Unfortunately, it is not possible to provide a quantitative assessment of the size of the reduction for any given estimate.

## Cost-Efficiency and Response Rates in the Phase 3 Sample

Table G presents the mean number of calls (computed on a segment basis) made on active cases in Phase 3. In the 1 month of Phase 3, an average of 4.08 calls were made on active screener cases. More effort was expended per case on segments with large numbers of active cases than in segments with small numbers (4.25 calls per case for high number of incomplete active cases v. 2.96 for low). Ideally, the level of effort would be more uniform across segments with different numbers of active cases. Such uniformity would produce more equal response rates across the strata.

Table H shows the mean number of calls on cases where main interviews were sought in Phase 3. There is some tendency for more effort in the segments with low propensities (for example, 5.26 calls per active case in the low-propensity group compared with 4.30 in the high-propensity group), again reflecting individual interviewer decisions on how to obtain interviews.

Table J presents overall average segment screener response rates (number of interviews/number of eligibles) for the Phase 3 cases. These segment screener response rates are an average

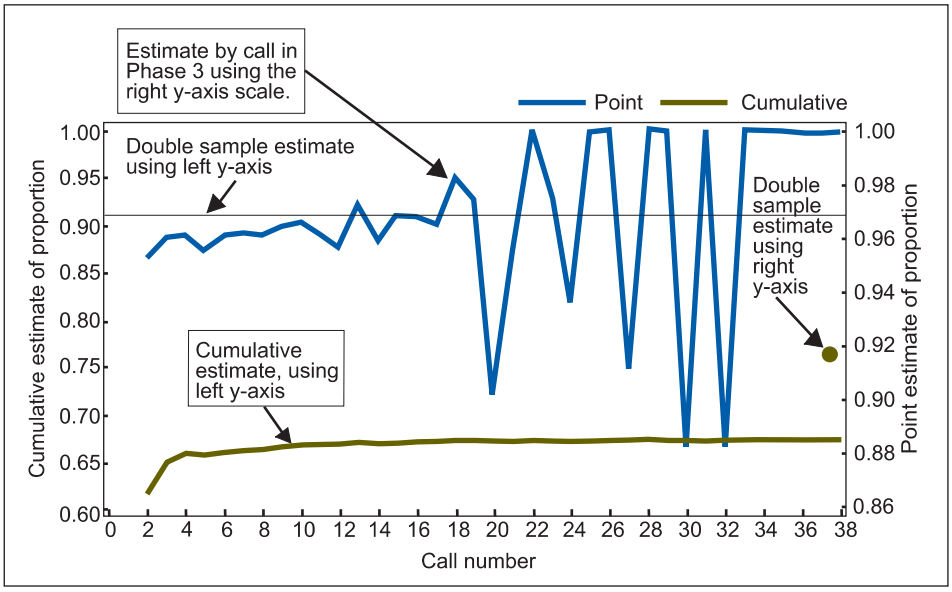


**Figure 8. Estimated proportion of females who have ever had sexual intercourse in Cycle 6 of the National Survey of Family Growth: Cumulative estimate, estimate by call number, and estimate from Phase 3 interviewing**

over the Phase 3 segments of the ratio of interviews to sample cases within the segments. The overall mean screener response rate per segment was 51 percent, a rather remarkable rate given the fact that the sample had been worked for 12 months in Phases 1 and 2. The response rates range from 0 to 89 percent for the active cases for which screeners were sampled.

Ideally, the response rates would be constant over all 16 cells of the Phase 3 design. There is much greater stability of these response rates at the margins, reflecting greater number of segments on which the means are based. Across the four strata of number of incomplete active cases, the response rates range from 49 to 55 percent with no trend by number of active cases. For the propensity strata, the range is from 49 to 53 percent, again with no trend by magnitude of propensities. This distribution is desirable, given the value of balancing representation of the Phase 3 sample over the 16 strata.

Table K presents the mean per segment response rates for the main interview cases. Across the 16 cells the variation in response rates is from 0 to 65 percent. The marginal variation, again, is much smaller. For strata defined by active case count, the range is from 42 to 54 percent, without a trend across strata. For strata defined by total propensities, the range is from 35 to 54 percent, with a trend toward higher response rates in the segments with higher propensities. This is undesirable, but reflects the difficulty of obtaining the main interviews after a prior refusal.

Tables L and M display the segment mean of the ratio of interviews to number of calls, an indirect indicator of the difficulty of obtaining an interview. The larger the ratio, the less the effort required to obtain interviews. It was expected that the smallest ratios would lie in the low-propensity strata.

Table L presents the results for the screener interviewing efforts. Here, the mean ratios range from 0.04 to 0.59 across the 16 cells. These correspond to an interview every 1.7 calls (that is, 1/0.59) to an interview every 25 calls. Looking at the margins of the table, there is an unexpected tendency for the segments with low propensities to have higher rates of interviews per call: from 0.35 in the low-propensity stratum to 0.54 for the high. The segments with low numbers of active cases also show higher efficiencies: from 0.42 for low number of incomplete active cases to 0.54 for high.

Table M shows interview per call rate for the main interview cases. Here, the efficiencies range from an interview every four calls (1/0.25 in the low propensity and low number of incomplete active cases) to zero. The next highest value is 1 in every 11 calls (1/0.09 in several cells). The higher efficiencies are found in the high-propensity strata (0.16 and 0.17 in the high and medium-high strata compared with 0.13 and 0.14 in the medium-low and low strata), as expected.

In the segments with fewer active cases, higher efficiencies were also observed (0.20 in the low number of cases stratum). This finding indicates that the survey climate in strata with fewer active cases was easier. There would be a tendency to find in segments with fewer active cases a better climate since the

**Table G. Mean number of calls per active screener case by number of active cases and total propensity strata in Phase 3: National Survey of Family Growth, Cycle 6**

| Estimated total propensity to respond for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 4.04 | 4.30 | 3.03 | 1.00 | 3.96 |
| Medium-high (next highest ¼) | 4.49 | 4.46 | 3.69 | 3.92 | 4.33 |
| Medium-low (next highest ¼) | 4.00 | 4.37 | 3.91 | 1.92 | 3.68 |
| Low (smallest propensities) | 7.06 | 4.22 | 3.00 | 5.11 | 5.02 |
| Total | 4.25 | 4.38 | 3.48 | 2.96 | 4.08 |

NOTE: Average over segments of the ratio of calls to segment sampled units not yet screened to number of cases screened during the third phase.

**Table H. Mean number of calls per active main interview case by number of active cases and estimated total propensity of cases in the segment to respond on the next visit: Phase 3 of the National Survey of Family Growth, Cycle 6**

| Total propensity for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 4.40 | 4.37 | 3.62 | 4.72 | 4.30 |
| Medium-high (next highest ¼) | 4.53 | 4.82 | 4.13 | 3.81 | 4.53 |
| Medium-low (next highest ¼) | 4.74 | 5.21 | 3.52 | 3.26 | 4.30 |
| Low (smallest propensities) | 5.44 | 5.11 | 7.11 | 4.00 | 5.26 |
| Total | 4.50 | 4.74 | 4.07 | 3.79 | 4.44 |

NOTE: Average over segments of the ratio of calls to sampled respondents to number of main survey respondents interviewed during the third phase.

**Table J. Mean screener response rates for segments in Phase 3 by number of active cases and total propensity strata: National Survey of Family Growth, Cycle 6**

| Total propensity for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 0.52 | 0.42 | 0.59 | 0.00 | 0.50 |
| Medium-high (next highest ¼) | 0.51 | 0.51 | 0.53 | 0.83 | 0.53 |
| Medium-low (next highest ¼) | 0.56 | 0.48 | 0.41 | 0.54 | 0.49 |
| Low (smallest propensities) | 0.28 | 0.89 | 0.50 | 0.33 | 0.52 |
| Total | 0.51 | 0.49 | 0.52 | 0.55 | 0.51 |

0.0 Quantity more than zero but less than 0.05.

NOTE: Average over segments of the ratio of completed screeners in third phase to number of cases not yet screened at beginning of the third phase.

lower number of incomplete active cases itself indicates that Phases 1 and 2 effort was able to complete more cases.

These results indicate that the Phase 3 stratification was more highly effective for the main than for the screener interviews. In retrospect, the effect of blending the screener and main response propensity models to form the Phase 3 strata led to this result. Most of the Phase 3 cases were main interview nonresponse, a set of cases where there was a stronger relationship between predictors and propensity to respond to the main interview. Using the ‘‘blended’’ model forced a better prediction for main than for screener interviews.

## Responsive Design

In previous cycles of the NSFG, as in many other surveys, the survey managers attempted to keep track of interviewer effort and field costs using a paper and pencil system to record each interviewer’s hours worked and other expenses. One problem with these systems was that it was often 3–6 weeks before their results were available. As a result, data on survey costs and effort lagged far behind data on completed cases and response rates, and there was little or no information that was specific enough to provide advice to interviewers that was based on detailed knowledge of their work. As a result, the NSFG contract required that the contractor have a system that could provide data on field costs and response rates that were more timely and thus more useful for controlling costs and managing fieldwork.

The ISR system for monitoring fieldwork, called SurveyTrak, required the interviewers to record simple characteristics of neighborhoods (sample segments), sampled households, and statements made by respondents to the interviewers during attempts to contact them for an interview. The contractor collected and analyzed these data during fieldwork, to provide data to allocate interviewer labor during the study. This information was used throughout fieldwork: in Phase 1, it was used to make estimates of interviewer productivity and survey costs; in Phase 2, it was used to help survey managers to allocate direct interviewer labor; and in Phase 3, it was used to sort the remaining (nonrespondent) cases to decide where to put interviewer labor to get the most cost-efficient return. This approach to survey management is called ‘‘Responsive Design.’’

A variety of statistical models were used to evaluate the progress of field work and to decide where to allocate interviewer effort in the final stages of interviewing:

1. A logistic regression model predicting the probability of a completed main interview, using as predictor variables the number of hours interviewers spent traveling, hours doing administrative tasks, hours making screener visits, and hours making main interview visits.
2. A linear regression model predicting interviewer hours needed to obtain a completed main interview, separately for each of the 18 demographic subgroups, using such predictors as total number of calls made to a case, an indicator for whether a contact

**Table K. Mean main interview response rates for segments during Phase 3 by number of active cases at the beginning of Phase 3, and the total propensity of cases in the segment to respond on the next visit: National Survey of Family Growth, Cycle 6**

| Total propensity for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 0.54 | 0.50 | 0.64 | 0.50 | 0.54 |
| Medium-high (next highest ¼) | 0.56 | 0.57 | 0.51 | 0.31 | 0.53 |
| Medium-low (next highest ¼) | 0.51 | 0.35 | 0.35 | 0.57 | 0.43 |
| Low (smallest propensities) | 0.33 | 0.00 | 0.65 | 0.33 | 0.35 |
| Total | 0.54 | 0.49 | 0.53 | 0.42 | 0.51 |

0.0 Quantity more than zero but less than 0.05.

NOTE: Average over segments of the ratio of main interviews in third phase to number of cases eligible for the main interview in the third phase.

**Table L. Mean ratio of number of screener interviews to number of total calls on screener cases by number of active cases in the segment, and by the estimated total propensity of cases in the segment to respond on the next interviewer visit: National Survey of Family Growth, Cycle 6**

| Total propensity for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 0.15 | 0.14 | 0.29 | 0.00 | 0.16 |
| Medium-high (next highest ¼) | 0.13 | 0.18 | 0.20 | 0.59 | 0.18 |
| Medium-low (next highest ¼) | 0.13 | 0.14 | 0.11 | 0.37 | 0.17 |
| Low (smallest propensities) | 0.04 | 0.47 | 0.50 | 0.04 | 0.24 |
| Total | 0.14 | 0.17 | 0.22 | 0.34 | 0.18 |

0.0 Quantity more than zero but less than 0.05.

NOTE: Average over segments of the ratio of completed screeners in third phase to total number of screener calls during the third phase.

had been made, and other operational variables.

3. Using the logistic and linear models in Phase 1 and Phase 2, a model to forecast the final number of interviews under different allocations of staff.

These models were used in conjunction with statistical tracking (9) to provide more efficient data collection.

In addition, simple statistics were tracked to identify key lifecycle changes in the data collection. For example, figure 9 shows the cumulative ratio of total interviewer hours worked to total main interviews collected. It has the pattern common to surveys involving screening interviews. The first weeks of Phase 1, especially in April 2002, based on the quarter-sample, were devoted mainly to identifying eligible sample persons and relatively few main interviews were obtained. This produced a higher average number of hours of labor per interview. Then there was a decline and a rapid increase toward the end of Phase 1 (the third week of June), when interviewers were mostly making repeated visits on those who were difficult to contact or those reluctant to be interviewed.

In the beginning of July 2002, new sample areas were opened, and new interviewers began to work in them, so that all 121 PSUs were being worked. The number of hours per interview rose again as the new cases were contacted and screened. Hours per interview gradually declined and reached a plateau of about 10.9 hours per interview in the fall months, and then began increasing again as the interviewers completed the easy-to-contact and easy-to-interview cases and focused on the more difficult

**Table M. Mean ratios of number of main interviews to number of total calls on screener cases by number of active cases and total propensity strata: National Survey of Family Growth, Cycle 6**

| Total propensity for incomplete active cases (Quartiles) | Number of incomplete active cases at end of Phase 2 (Quartiles) | | | | |
|---|---|---|---|---|---|
| | High | Medium-high | Medium-low | Low | Total |
| High (top ¼ of propensities) | 0.16 | 0.14 | 0.22 | 0.10 | 0.16 |
| Medium-high (next highest ¼) | 0.16 | 0.15 | 0.22 | 0.19 | 0.17 |
| Medium-low (next highest ¼) | 0.11 | 0.09 | 0.11 | 0.23 | 0.13 |
| Low (smallest propensities) | 0.13 | 0.00 | 0.11 | 0.25 | 0.14 |
| Total | 0.16 | 0.13 | 0.18 | 0.20 | 0.16 |

0.0 Quantity more than zero but less than 0.05.

NOTE: Average over segments of the ratio of completed main interviews in third phase to total number of main interview calls during the third phase.

cases. The February 1, 2003, Phase 3 introduction prevented the rapid increase in hours per interview seen in Phase 1 by implementing more efficient protocols on a subsample of the remaining nonrespondent cases.

## Sample Size

The use of a responsive design for Cycle 6 effectively managed the survey to achieve the maximum number of completed interviews while remaining within the fixed budget constraints of the survey. The responsive design approach continually examined expected and achieved eligibility and response rates throughout the data collection period. As differences were observed, sample release was altered to obtain the largest total number of completed main interviews, and a minimum number of completed interviews with 15–19-year-olds and black and Hispanic persons. A comparison of initial target sample size to the final achieved results is less meaningful in Cycle 6 under a responsive design.

Table N presents the number of completed interviews by 18 subclasses formed by age group, gender, and race and ethnicity group. (The race and ethnicity grouping is that used in the sample design: black, Hispanic, and other, where ''other'' includes non-Hispanic white, Asian, and American Indian respondents; see ''race'' in Appendix I, the ''Definition of Terms''). Sample size is an often used indicator of the reliability of estimates obtained from a survey. In Cycle 6, the sample size was only 145 for black males 20–24 years of age. Teen sample sizes for both genders and all race and ethnicity groups are larger, ranging from 205 for black males 15–19 years to 681 for (white and) other race males 15–19 years. The most reliable estimates will be obtained for (white and) other race females 25–44 years of age with a sample size of 2,995, followed by (white and) other race males 25–44 years with a sample size of 1,616.

In a stratified multistage sample survey like the NSFG Cycle 6, sample size is not the only feature of the sample that affects the reliability of estimates. Because the sample is clustered, and because the estimates must be weighted, the estimated levels of precision can be expected to be lower when these factors are accounted for. The user of the microdata must be aware that standard statistical estimation software does not properly account for these factors in estimation of reliability (that is, variances). Procedures are given later in this report on how to use statistical estimation software that will properly account for the clustering, stratification, and weighting employed in the NSFG Cycle 6 sample.

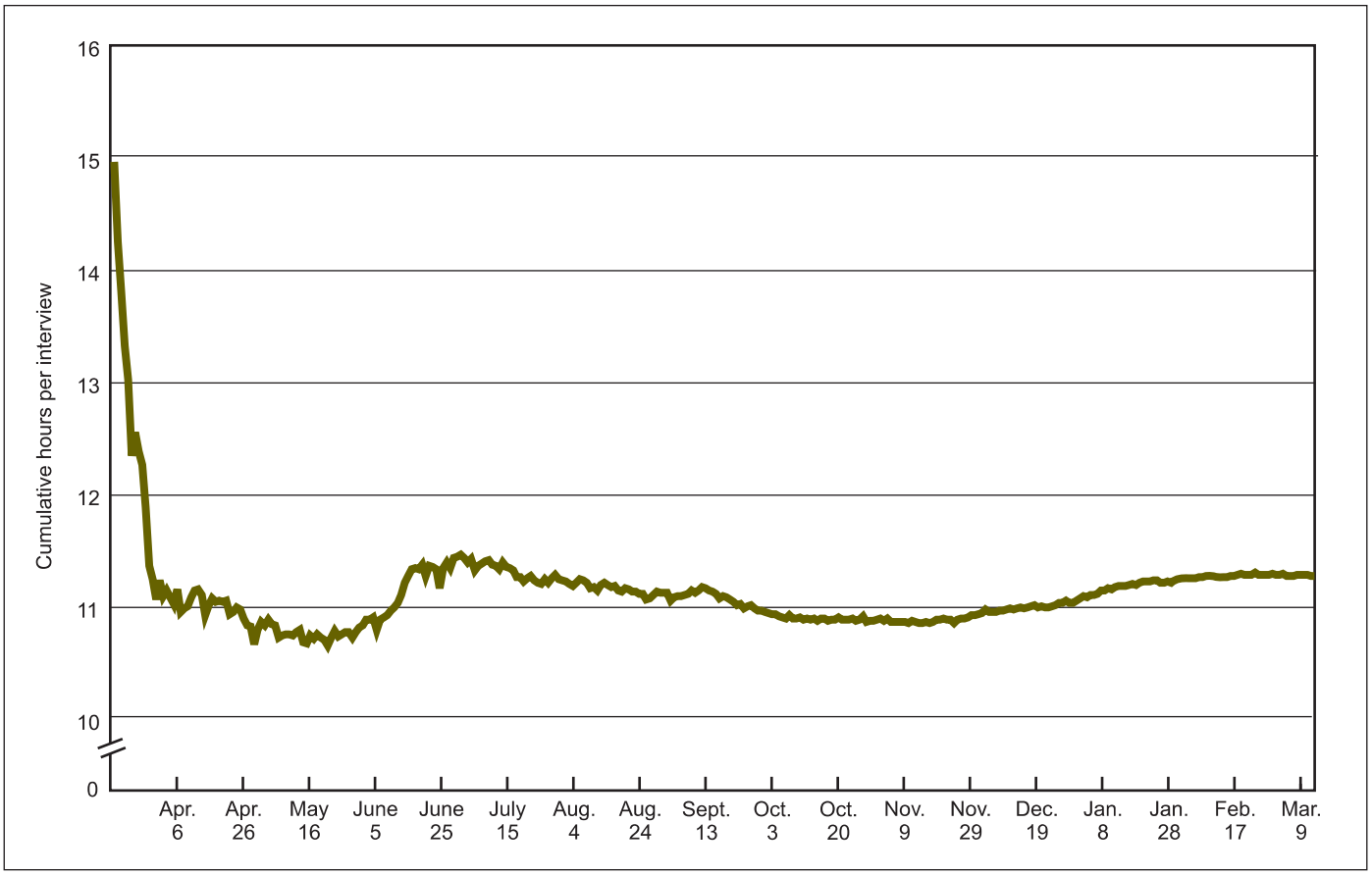


**Figure 9. Cumulative ratio of interviewer hours to number of main interviews completed by day of the data collection period: National Survey of Family Growth Cycle 6, 2002**

**Table N. Number of completed interviews by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6, 2002**

| Sex and age | Race or ethnicity | | | |
|---|---|---|---|---|
| | Total | Black | Hispanic | Other |
| Both sexes | | | | |
| Total | 12,571 | 2,460 | 2,712 | 7,399 |
| 15–19 years | 2,271 | 447 | 466 | 1,358 |
| 20–24 years | 2,301 | 363 | 508 | 1,430 |
| 25–44 years | 7,999 | 1,650 | 1,738 | 4,611 |
| Female | | | | |
| Total | 7,643 | 1,530 | 1,589 | 4,524 |
| 15–19 years | 1,150 | 242 | 231 | 677 |
| 20–24 years | 1,363 | 218 | 293 | 852 |
| 25–44 years | 5,130 | 1,070 | 1,065 | 2,995 |
| Male | | | | |
| Total | 4,928 | 930 | 1,123 | 2,875 |
| 15–19 years | 1,121 | 205 | 235 | 681 |
| 20–24 years | 938 | 145 | 215 | 578 |
| 25–44 years | 2,869 | 580 | 673 | 1,616 |

# Weighting Procedures

## Summary

NSFG Cycle 6 is designed to estimate the number of men and women with particular characteristics in the U.S. household population. For example, the NSFG can provide estimates of the number of women who have ever had sexual intercourse, not only for the total population but also for age groups, such as teenagers 15–19 years of age, men ages 20–24 years, or Hispanic women ages 25–44 years.

The NSFG deliberately oversampled specified subgroups of the population. It is also subject to nonresponse and noncoverage error. Sampling weights have been assigned to every person in the public-use data file. These weights must be used to compensate for unequal probabilities of selection, and for underrepresentation of some subgroups due to nonresponse or noncoverage. These weights are designed to produce estimates from the sample that correct for oversampling, nonresponse, and noncoverage.

The weights were constructed in four steps. In the first step, a set of weights was constructed to account for various deliberate over- and undersampling features of the Cycle 6 sample design. These weights were combined into a single ''base weight'' to represent the probabilities of selection.

In the second step, an extensive set of nonresponse models was developed to predict the probability of response to the screener for all households that contained eligible persons and for all selected eligible persons within successfully screened households. These models used predictors obtained from the housing unit listing process, interviewer observations collected at all sample households from all contacts with the household, household call records, household composition, and a few additional household and selected person characteristics (for models on eligible persons). The inverse of the predicted probabilities were used as adjustment factors applied to the base weights.

Third, nonresponse adjusted base weights were cumulated within each of 36 groups—6 age categories, 2 categories of gender, and 3 race and ethnicity subgroups. The cumulated weights were compared with independent Census Bureau population projection estimates for each cell, plus counts of the military personnel in each cell who were living off-base obtained from the U.S. Department of Defense. Poststratification weights were computed for each cell as the ratio of Census plus Defense counts to the cumulated sample weights in each cell. The poststratification factors were used to adjust the individual nonresponse adjusted base weights.

Fourth, the poststratified weights were inspected to detect extremely large or small weight values. No weight values were deemed to be extremely small. A total of 49 large weight values were ''trimmed'' to reduce their impact on weighted estimates. The post-stratification factors for the respective cells of each trimmed weight were recomputed after trimming. A more detailed account of the weighting procedures follows.

## Inverse Probability Selection Weighting

The NSFG Cycle 6 sample management system preloaded required information to calculate the probability of selection of each sample case. The inverse of the probability of selection of each case is the base weight adjusting for unequal chances of selection.

There are several components to the probability of selection and the base weight. Each case was selected through four stages of sampling: primary, segment, housing unit, and person within selected household. The primary stage of selection involved two overlapping samples of primary sampling units: a national sample and a Hispanic sample. In addition, the Phase 3 sample selected a subset of cases for intensive followup for the last month of fieldwork.

## Primary Sampling Units

In the national sample, the land area of the United States, including the District of Columbia, was divided into 3,141 counties and county equivalent units (see sample design description for details). These units, in turn, were combined to form a set of 2,402 primary sampling units (PSUs), which were then grouped into 110 strata (including Alaska and Hawaii) for the first-stage selection of the NSFG Cycle 6 national sample and 11 additional strata for the first-stage sample of the Hispanic sample. The combination of the national and Hispanic samples yielded 121 NSFG Cycle 6 PSUs.

The probability of selection of each PSU in each sample was computed, accounting for the overlap in selections, at the time of selection and stored for later use. All 121 PSUs were subsampled in NSFG Cycle 6.

## Segment Selection

Within each sample PSU, the land area was further divided into blocks, as defined by the 2000 census. The blocks were divided into four domains within each PSU:

1. Nonminority
2. More than 10 percent black population
3. More than 10 percent Hispanic population
4. More than 10 percent black and more than 10 percent Hispanic population

Blocks were selected within each domain with probabilities proportionate to the estimated number of households in the area in the year 2000.

To improve data collection efficiency by reducing travel costs, census blocks falling below a minimum number of estimated households were linked with other geographically contiguous blocks. In urban areas, the minimum size was 75 estimated households, while in rural areas the minimum was 50. The final units consisted of entire census blocks or set of linked census blocks and are referred to as segments. There were a total of 1,414 segments selected across the 4 domains and the 121 PSUs.

In all PSUs, segments in domains 3 and 4 were chosen with 2.1 times higher probabilities of selection than those in domain 1. Those in domain 2 were chosen with a 1.8 times higher rate. Probabilities of selection were proportionate to the estimated number of households in the segment. The probability of selection of each block was computed at the time of selection and stored for later use.

## Housing Unit Selection

Not all sample segments were subsequently selected for the NSFG sample. A larger number of segments were selected initially to have available a sufficient number of housing units in each segment to achieve, if needed, higher target sample size in the responsive sample design.

Segments were subsampled in each of the three phases of the sample release. Across Phases 1 and 2 a total of 783 segments were selected for the NSFG Cycle 6. Probabilities of selection for the Phases 1 and 2 segment sampling were computed for each segment and stored for later use.

Each segment was subsequently visited by an ISR interviewer trained to list all housing units in the segment. Sample listings were returned to Ann Arbor, where they were keyed and stored in a file containing ‘‘sample lines,’’ an entry for a listed housing unit.

Sample lines were selected to achieve a large initial target sample size of more than 72,000 lines. Each sample line (listed housing unit) was assigned to 1 of 24 replicate subsamples within each segment. Phases 1 and 2 sampling used replicate samples to achieve target sample sizes for each phase of selection to be released for interviewers hired from the PSU. The subsampling probabilities of selection for housing units were recorded in internal memoranda to be used later in weight computation.

## Person Selection

Once sample lines (housing units) were selected, they were released through the ISR sample management system to interviewers in each PSU. Interviewers visited sample housing units and determined if any eligible persons resided there. If eligible persons did reside in the housing unit, the interviewers completed a household roster where they recorded the age, gender, and race and ethnicity of each eligible person.

In the interview, a random selection was made of one eligible person per household. The chances of selection within the household were varied to increase the chances of selection of teens (ages 15–19), women, and black and Hispanic persons. The household roster and chances of selection were recorded in the household interview.

## Nonresponse Sampling

A final component of sample selection was introduced in Phase 3 sampling in the final month of data collection. A sample was chosen for more intensive data collection effort. At the last month of data collection, most of the 783 segments selected for the sample had sample lines that were not completed. Segments containing sample lines that had not reached a final disposition 1 month before the end of data collection were divided into 16 strata on the basis of the number of active cases remaining and predicted probabilities of obtaining a completed interview cumulated across all remaining cases in the segment.

A Phase 3 selection of segments from each of the 16 groups was made, with higher chances of selection assigned to those segments with higher expected numbers of completed interviews and higher numbers of cases remaining. The chances of selection were thus varied for the selection of second-phase segments released for data collection in the last month of the survey period. These probabilities of selection were retained for subsequent weighting.

## Probability of Selection and Weight

The probability of selection of each sample person can thus be computed using the probabilities of selection for PSUs, segments

(including segment replicate), sample line and sample line replicate, within-household selection, and Phase 3 subsampling of replicates.

Let $M_{h\alpha}$ denote the size measure of the $\alpha^{th}$ PSU in PSU stratum $h$. For national sample PSUs, the measure of size was the number of occupied households in the PSU in 1990, while for the Hispanic sample, the measure of size was the 2000 Hispanic population.

Let $M_{h\alpha\beta}$ denote the size measure for the $\beta^{th}$ segment in the $(h\alpha)^{th}$ PSU. The size measure for each segment was an estimated number of households obtained by dividing the census population for the blocks in the segment by the average number of persons per household obtained from the March 1999 CPS. Also, let $c_{dh\alpha}$ denote the desired number of sample lines for the $d^{th}$ domain in the $(h\alpha)^{th}$ PSU.

Let $\pi_p$ denote the probability of selection for the national sample quarter sample PSU replicate, $\pi_{s,h\alpha\beta}$ denote the sample replicate selection probability for the $(h\alpha\beta)^{th}$ segment (which can take on values of 1/3, 2/3, or 1, depending on the subsampling rate employed in the PSU and domain), and let $\pi_{hu,h\alpha\beta}$ denote the replicate selection probability for the housing unit or sample line in the $(h\alpha\beta)^{th}$ sample segment. Finally, let $\pi_{2,h\alpha\beta}$ denote the Phase 3 selection probability for the $(h\alpha\beta)^{th}$ sample segment.

The probability of selection of the $(h\alpha\beta\gamma)^{th}$ eligible person (where $\gamma$ denotes the $\gamma^{th}$ sample person in the $(h\alpha\beta)^{th}$ segment) can be computed as

$$\pi_{h\alpha\beta\gamma} = \left(\frac{M_{h\alpha}}{\sum_{\alpha=1}^{a_h} M_{h\alpha}}\right) \bullet \pi_{p,h\alpha} \bullet \left(\frac{M_{h\alpha\beta}}{\sum_{\beta=1}^{b_{h\alpha}} M_{h\alpha\beta}}\right) \bullet \pi_{s,h\alpha\beta} \bullet$$

$$\pi_{2,h\alpha\beta} \bullet \left(\frac{c_{dh\alpha}}{M_{h\alpha\beta}}\right) \bullet \pi_{hu,h\alpha\beta}$$

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

The base weights for the 12,571 completed interviews are summarized in table O for each of the target 18 subclasses. Large values of these weights were not trimmed in the weighting process.

Because of the large number of components, the base weights have substantial variation within subclasses. For example, the base weight values for female other race respondents ages 25–44 range from 215 to 137,124. The largest contributing factor to this very wide range is the subsampling within households described previously and summarized in figures 4 and 5. In order to achieve a large number of completed interviews with teens 15–19 years of age, a substantial oversample within households was necessary. The oversampling of teens led to a substantial undersampling of 25–44 year olds and much larger weights for the 25–44 year olds in households where teens lived. At the same time, 25–44 year olds living in households without teens had much higher chances of selection and had much lower weights.

This variation in base weights due to the over- and undersampling within households has the potential to increase the variance of estimates. A summary

**Table O. Mean, minimum, and maximum untrimmed base weight, and potential increase in variance due to weighting (1+*L*), by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6**

| Sex, race or ethnicity, and age | Mean weight | Minimum weight | Maximum weight | Increase in variance (1+*L*) |
|---|---|---|---|---|
| **Male** | | | | |
| Black: | | | | |
| 15–19 | 4,556.55 | 1167.49 | 31,868.62 | 1.6853 |
| 20–24 | 4,936.14 | 405.35 | 51,268.90 | 2.8206 |
| 25–44 | 4,676.89 | 366.60 | 69,461.98 | 2.6211 |
| Hispanic: | | | | |
| 15–19 | 4,944.55 | 1167.57 | 23,432.59 | 1.4150 |
| 20–24 | 4,857.30 | 435.19 | 31,967.17 | 1.7063 |
| 25–44 | 5,224.50 | 1167.44 | 37,341.97 | 1.7702 |
| Other: | | | | |
| 15–19 | 4,754.88 | 1167.44 | 20,649.67 | 1.3072 |
| 20–24 | 5,915.56 | 1109.84 | 29,392.20 | 1.5592 |
| 25–44 | 7,309.55 | 215.26 | 137,124.02 | 2.6746 |
| **Female** | | | | |
| Black: | | | | |
| 15–19 | 3,803.41 | 766.42 | 13,893.39 | 1.3681 |
| 20–24 | 3,537.10 | 766.42 | 35,394.72 | 1.8366 |
| 25–44 | 3,126.23 | 366.60 | 21,375.42 | 1.5419 |
| Hispanic: | | | | |
| 15–19 | 4,488.10 | 1167.57 | 18,681.57 | 1.3745 |
| 20–24 | 3,772.34 | 1167.44 | 12,118.88 | 1.2802 |
| 25–44 | 3,559.23 | 366.60 | 22,426.69 | 1.4794 |
| Other: | | | | |
| 15–19 | 4,696.32 | 205.10 | 18,183.37 | 1.2703 |
| 20–24 | 4,478.61 | 478.56 | 47,029.78 | 1.4232 |
| 25–44 | 4,631.78 | 64.58 | 99,707.83 | 1.7192 |

factor of the potential increase in variance, computed under the assumption that the weights are uncorrelated with the values of the characteristics being estimated, is estimated for each cell. The factor $1+L$ is a measure of the relative variance of the weights themselves, and is computed for the total sample as

$$1+L = \frac{12{,}571\left(\sum_{i=1}^{12{,}571} w_{1i}^2\right)}{\left(\sum_{i=1}^{12{,}571} w_{1i}\right)^2}$$

This factor is an approximation that indicates the relative effect of weighting across the cells. It is not an exact factor that can be attributed to each case in a cell.

Nonetheless, there are substantial potential increases in variance because of weighting across all cells. The largest values of the factor $1+L$ are for females, particularly black persons 20–24 and 25–44 years of age and other race and ethnicity ages 25–44 years. These largest increases can be directly attributed to the oversampling of 15–19-year-olds within households.

The full effects of these increases are reduced later in the weighting process through trimming of the largest final weight values. Further, sampling variance estimates presented in the last section of this report do not show large design effects that would in part be generated by the increase in variance due to weighting. Thus, the increases in variance shown in table O, and in subsequent tables in this section, should not be interpreted as actual direct effects on sampling variance estimates obtained by NSFG Cycle 6 analysts who properly account for the sample design in analysis.

## Nonresponse Adjustment

Nonresponse in NSFG Cycle 6 occurred at both screening to identify sample eligible persons in sample households and at the main interview among selected eligible persons. At the screener interview, nonresponse occurs at the time of initial contact when limited information is available about the nonresponding housing unit. The main interview nonresponse occurs any time after the conclusion of screening, when a sample person has been selected. The main interview nonresponse has additional information about household composition and race and ethnicity of the selected person, obtained during household screening.

Nonresponse adjustment for NSFG Cycle 6 was implemented under an assumption of missing at random (MAR). That is, within subgroups of sample units (housing units in the screener, selected eligible persons in the main interview), it is assumed that the nonrespondents are a random sample from all the units in the sample. A nonresponse weighting adjustment developed under this assumption can be computed as the inverse of an estimated response rate. This sample based weight represents an adjustment that, under the MAR assumption, may substitute for a probability of selection in the response process. Thus, as for unequal probability weighting, the inverse of the predicted probability of response serves as an adjustment factor.

There are many methods for developing estimated probabilities of selection under MAR. One is to divide the sample into weighting classes across which response rates are expected to vary, and across which the characteristics of sample persons are expected to vary. A more general form of the weighting class analysis is to estimate probabilities of response through some form of a logit or propensity model, as in a logistic regression. Separate models for screening and main interview nonresponse were developed.

Let $S_i$ denote a zero-one indicator for whether a sample address was successfully screened to determine whether eligible persons lived in the household, where 1 denotes successful screening. ($S_i$ is not defined for sample addresses that were not occupied.) The screening logistic regression model for all occupied sample addresses is

$$\lambda_{si} = Pr\,(S_i=1|X_i) = (1+exp\,(-X_i'\beta))^{-1}$$

where $X_i$ is a vector of predictor values for the $i$th occupied housing unit and $\beta$ is a vector of logistic regression coefficients. The coefficient values can be estimated through standard maximum likelihood methods, and used to predict the logit

$$\hat{\lambda}_{si} = (1+exp(X_i'\hat{\beta}))^{-1}$$

which, in turn, can be used to obtain a predicted probability of successful screening is

$$\hat{\pi}_{si} = \frac{exp(-\hat{\lambda}_{si})}{(1+exp(-\hat{\lambda}_{si}))}.$$

Similarly, let $R_i$ denote a 0–1 indicator for main interview response for the $i$th successfully screened occupied housing unit, equal to 1 when the selected eligible person has a completed interview, and 0 for selected eligible persons who did not complete an interview. ($R_i$ is not defined for sample addresses that are not occupied or were not successfully screened.) The main interview logistic regression model for selected eligible persons is

$$\lambda_{mi} = Pr(R_i = 1 \mid S_i = 1, Z_i) = (1+exp(-Z_i'\gamma))^{-1}$$

where $Z_i$ is a vector of predictor values for the $i$th selected eligible person and $\gamma$ is a vector of logistic regression coefficients. Standard maximum likelihood methods yield

$$\hat{\lambda}_{mi} = (1+exp\,(Z_i'\hat{\gamma}))^{-1}.$$

The predicted probability of successful screening

$$\hat{\pi}_{mi} = \frac{exp\,(-\hat{\lambda}_{mi})}{(1+exp\,(-\hat{\lambda}_{mi}))}.$$

That is, a model was developed for screening, and a second model was developed for the main interview given successful screening. The overall response propensity is then estimated for each housing unit based on the product of two predicted probabilities, $\hat{\pi}_{ri} = \hat{\pi}_{si} \times \hat{\pi}_{mi}$. This predicted probability is computed for all selected eligible persons.

A nonresponse weight was then computed as the inverse of the predicted probability, but only for completed interview cases. In particular,

$w_{ri} = \hat{\pi}_{ri}^{-1}$ if $R_i = 1$ and $S_i = 1$, and $w_{ri} = 0$ otherwise.

Unlike NSFG Cycle 5, which had the National Health Interview Survey data set for each sample person to use as potential predictors in response propensity models, Cycle 6 has a more limited set of geographic and operational variables to use as predictors for the screening and main interview models. Preliminary modeling of screener and conditional main interview response propensity was performed in the estimation of segment level expected completed interview counts for the Phase 3 sample selection (9). Those preliminary models employed a number of predictors:

1. Counts and rates for the segment from which the housing unit was selected, derived from 2000 census data for the blocks in the segment.
2. Data obtained from observations made at two levels for each housing unit: characteristics of the segment and housing unit recorded by the interviewer who listed the segment.
3. Respondent behavior recorded by the interviewer at each contact with anyone within the housing unit.
4. Operational measures, such as number of calls to a housing unit, number of calls to the sample person, and interviewer response rate.
5. For the main interview propensity model, data were drawn from the household roster and other data collected in the screening interview.

These sets of variables were used as predictors in the response propensity models for screener and main interview. The census variables, segment and housing unit observations, screener contact observations, and operational variables were examined in stepwise logistic regressions for the screening indicator variable. The same predictor variables plus main interview contact observations, household composition, and other variables available from the screening interview were examined in stepwise logistic regressions for the main interview response indicator $R_i$.

The predictors examined in the Phase 3 sample selection propensity models were re-examined for the screening and main interview response propensity models, again using stepwise logistic regression to identify a set significantly associated with response. The screener and main interview stepwise models yielded different sets of predictors, despite considerable overlap in predictors between the models. There are also theoretical considerations for choice of predictors between the two models. The development of each model will be described separately.

## Screener Response Propensity Model

The predictors in the screener propensity model and their estimated coefficients are shown in table P. Not only did all of these variables have significant association with response to screening, but also these variables represent a theoretically important set of predictors of response propensity. For example, the survey methodology literature (10,12) demonstrates that certain behaviors recorded in contact observations are highly correlated with response propensity. Four "ever ..." variables stand out among the predictors:

1. **everresist:** during any contact with the household, any respondent refused to participate in the survey
2. **everquest:** during contact with the household, any respondent asked a question about the content or objectives of the survey
3. **evercontact:** during any contact with the household, any respondent was contacted and provided information about the household
4. **everdelay:** during any contact with the households, any respondent made an appointment for an interview at another time

**Table P. Logistic Regression Model Predicting Response to the Screener, with 18 predictors (Phases 1–3): National Survey of Family Growth, Cycle 6**

| Parameter | Estimate | Standard error | Pr > ChiSq |
|---|---|---|---|
| Intercept | 0.7092 | 0.09020 | <.0001 |
| Everresist[1] | –2.7902 | 0.05150 | <.0001 |
| Everquest[2] | 0.9388 | 0.04420 | <.0001 |
| Release 1[3] | 1.3875 | 0.07780 | <.0001 |
| Release 2[4] | 1.0110 | 0.07470 | <.0001 |
| Release 3[5] | 0.8408 | 0.08300 | <.0001 |
| Evercontact[6] | 0.1998 | 0.05370 | 0.0002 |
| Callstocontact[7] | –0.0526 | 0.00631 | <.0001 |
| Everdelay[8] | –0.3625 | 0.04760 | <.0001 |
| Physimped[9] | –0.3383 | 0.05460 | <.0001 |
| Uninhab[10] | 0.2613 | 0.04820 | <.0001 |
| Bilingual[11] | 0.2944 | 0.05440 | <.0001 |
| RaceID[12] | 0.0265 | 0.01120 | 0.0178 |
| Lspanish[13] | 0.2887 | 0.05150 | <.0001 |
| Tenured[14] | 0.7675 | 0.04330 | <.0001 |
| Urban[15] | –0.1777 | 0.04510 | <.0001 |
| Manyunits[16] | –0.3801 | 0.04630 | <.0001 |
| Evepct[17] | 0.0157 | 0.00151 | <.0001 |

[1]Everresist: during any contact with the household, any respondent refused to participate in the survey.
[2]Everquest: during contact with the household, any respondent asked a question about the content or objectives of the survey.
[3]Release 1: indicator that sample line was released during Phase 1.
[4]Release 2: indicator that sample line was released during Phase 2.
[5]Release 3: indicator that sample line was released during Phase 3.
[6]Evercontact: during any contact with the household, any respondent was contacted and provided information about the household.
[7]Callstocontact: indicates the number of calls until contact was made.
[8]Everdelay: during any contact with the households, any respondent made an appointment for interview at another time.
[9]Physimped: indicator of whether or not there was a physical impediment to accessing the housing unit.
[10]Uninhab: during screening the housing unit was unoccupied.
[11]Bilingual: indicator that housing unit is bilingual.
[12]RaceID: during field listing of the segment, the listing staff identified the predominant race of households in the segment as black.
[13]Spanish: during field listing, the listing staff reported that there were Spanish speaking households in the segment.
[14]Tenured: during field listing, the listing staff reported that the household was a single-family dwelling.
[15]Urban: indicator that housing unit was in an urban area.
[16]Manyunits: indicator that housing unit was a multi-unit structure.
[17]Evepct: percentage of calls made to the household during evening hours after 6 PM local time.

All "ever" variables were derived from contact information. The signs of the coefficients for each of the "ever" variables in table P are consistent with the theoretical and empirical literature. In particular, resistance, delays, and failure to contact persons in the housing unit during a visit reduced response propensity, while informant questions (everquest) asked during contact increased response propensity.

Other variables in the model had substantial associations with screener response as well:

1. When the sample line was released (during Phase 1 in March (*release1*), Phase 2 in June (*release2*), or Phase 2 in September (*release3*)) indicates that later release yielded lower response propensity.
2. Physical impediments to access (*physimped*), multiunit structures (*manyunits*), and urban location (*urban*) were associated with lower response rates.
3. More calls to the housing unit to complete the screening interview were associated with lower response propensity (*callstocontact*).
4. Housing units in bilingual (*bilingual*), Spanish speaking (*lspanish*), and nonblack (*raceID*) segments, and owner occupied housing units (*tenured*) had higher response propensities.
5. Higher percentages of evening calls (*evepct*) were associated with increased response propensity.

Predicted probabilities of response ranged from near 0.05 to 0.98. This substantial variation in predicted probabilities generates considerable variation in response propensity weights. A common practice in survey estimation is to reduce this variation by grouping predicted values into classes, and then using some middle value to represent the entire group of predicted values. This method effectively imposes a constraint on the underlying propensity model. The predicted probabilities were grouped by deciles of the predicted probabilities. All completed screener cases in the lowest predicted probability decile were assigned the value of the median of the predicted probabilities in the decile. The remaining 90 percent of the values were not changed. This response propensity revision reduced the variation to a range from 0.39 to 0.98. The main interview nonresponse weight was computed as the inverse of the predicted probability, or the inverse of the median of the lowest decile.

## Main Interview Response Propensity Model

The main interview response propensity model development was more complicated than the screener propensity model because a larger number of predictors were available, and a broader theoretical framework could be used for weight development. The stepwise logistic regression indicated that age, gender, and race and ethnicity were not strong predictors of response propensity for the main interview model. However, all are known to be strong predictors of the values of fertility and other characteristics observed in the interview.

Theoretically, there are two general approaches to developing weights to adjust for nonresponse. The response propensity approach seeks to find a set of subjects whose predicted probabilities of response are identical, or nearly identical, and then assign to the respondents in the set a common weight that compensates for the nonrespondents in the set. Because there is no variation among the cases in the set on the propensity to respond, there is no association between characteristics of interest and response propensity. That is, nonresponse is independent of the characteristics of interest. At the same time, individuals with different propensities (that is, in different sets) who also have different values of the characteristic will generate association between propensity and substantive measures. The inverse propensity weights thus allow the association between propensity and substantive values to be used to adjust the estimates.

A second approach is the predictive mean approach. If a group of sample persons who have the same or similar propensities have the same value of a substantive measure, then nonresponse is independent of the substantive measure. Models that predict substantive measures could be used to group respondents together based on predicted values of substantive variables. Inverse propensity weights within these groups will then allow the association between propensity and substantive measures to be used to adjust estimates.

The predictive mean approach is difficult to implement in practice because it is seldom the case that available predictors of substantive measures are present. However, in NSFG Cycle 6, age, gender, and race and ethnicity are known for all selected sample persons, and these are important predictors of substantive measures. Thus, while the response propensity model stepwise regression screening did not select these predictive variables, it is important to include them in some form in the propensity models. That is, a combination of response propensity and predictive mean models can be used for the main interview nonresponse adjustment.

Three different types of models were examined for main interview propensity. In the first, all statistically significant predictors in the stepwise regression were included, as well as age, gender, and race and ethnicity indicators. The second model included the same predictors, but added interactions between age and gender and the "Ever . . ." variables from the contact observation data that were highly associated again with response.

A third model employed a reduced set of only 6 predictors in each of 18 subclasses defined by age, gender, and race and ethnicity. These 18 separate models were based on samples that ranged in size from approximately 200 for black male teens to more than 4,000 for white and other race females 25–44 years of age. These sizes were too small for nine of the subclasses where quasi-complete separation prevented estimation of some model coefficients. As a result, the 18 subclasses were collapsed into 9 subgroups, each with substantially larger sample sizes, from approximately 1,200 to more than 4,000:

1. Male 15–19 years
2. Female 15–19 years
3. Male 20–24 years
4. Female 20–24 years

5. Male black or Hispanic 25–44 years
6. Male white and other race 25–44 years
7. Female black 25–44
8. Female Hispanic 25–44
9. Female white and other race 25–44

For each subclass model, the same six predictors are used in each subgroup model:

1. ever resisted interview
2. ever asked questions
3. ever delayed interview
4. ever failed to contact person in housing unit
5. screening interview completed in Spanish
6. single-person household (table Q)

These six predictors have the strongest association with main interview response propensity in the stepwise logistic regression selection. The three models each showed similar levels of change in the logarithm of the likelihood from the "null model" (containing only an intercept as the predictor to the model). Examination of the properties of the estimators and predicted values across the three models revealed no particular features of one model that would favor one over the other. A choice was made to use the nine subgroup model to generate predicted response propensities. It combined most completely the strongest response propensity predictors with the most important predictors available for a predictive mean approach.

The estimated coefficients for this model are shown in table Q. For the most part, the six predictors are associated with higher response propensity in each of the nine subclasses in the same way. More resistance (*everresist*), delayed interviewing (*everdelay*), and Spanish language of interview (*SCR_LANG*) are associated with lower response propensity. Questions (*everquest*) or contact at each call (*evercontact*) and single-person households (*SCR_SINGLEHH*) are associated with higher response propensities, although single-person households are in two subgroups only marginally significant. Estimated coefficients do vary across subclasses; the hypothesis of subgroup coefficient equality was not examined.

As for the screener interview propensity model, the main interview response propensity weights showed substantial variation, based on predicted probabilities that ranged from 0.05 to 0.98. The variation was reduced by the same technique applied to the screener response propensity model probabilities: cases in the lowest decile of predicted probabilities were assigned the value of the median of the predicted probabilities in the decile. The remaining 90 percent of the values were not changed.

The main interview nonresponse weight $w_{mi}$ was computed as the inverse of the predicted probability or the inverse of the median of the lowest decile.

### Combined Nonresponse Weight

As a final step in the construction of the nonresponse adjustment weight, the screener nonresponse weight $w_{si}$ was multiplied by the main interview nonresponse adjustment to obtain the final nonresponse adjusted weight, $w_{2i} = w_{si} \times w_{mi}$. No trimming of this final weight value was needed.

Table R summarizes the distribution of the combined nonresponse weights within each of 18 subclasses of substantive importance. The range of nonresponse weights is limited, and the potential increase in variance due to weighting is much smaller than for the base weight. When these adjustments are applied to the base weights to yield the nonresponse adjusted base weights, $w_{3i} = w_{2i} \times w_{1i}$, the distributions observed in table O for the base weight were changed only slightly.

## Poststratification

The final stage in the weighting process is the adjustment of weighted sample values to outside distributions. Preliminary study indicated that there are few external data to which the NSFG Cycle 6 data could be benchmarked in a calibration approach.

A calibration to external population counts by age, gender, and race and ethnicity was implemented. Poststratification adjusts the nonresponse adjusted weight $w_{3i}$ to the civilian noninstitutionalized population projections for June 2002 provided by the Census Bureau for 36 age, gender, and race ethnicity cells (see table S). (The Census Bureau projections shown in table S will not agree with published projections on the census Web site since the published projections undergo revision regularly, and change over time.) The poststratification weight $w_{4i}$ is the ratio of the combined Census and Defense counts to the sum of the nonresponse adjusted weight $w_{3i}$ in each cell. A preliminary final weight $w_{5i} = w_{4i} \times w_{3i}$ was computed for each of the 12,571 completed interview cases.

## Weight Trimming

The distribution of the preliminary final weight was examined in each of the primary 18 subclasses as well as by groups formed by age, gender, and race and ethnicity, and two-way combinations of these variables. Potential increases in variance greater than 2.0 were observed in 4 of the 18 subclasses. After careful investigation, it was recognized that considerable reduction of the variability in the weights could be achieved by an arbitrary reduction of a few extremely large weight values. These kinds of *ad hoc* weight-trimming procedures are used on occasion in many surveys, but there is little theoretical justification for the procedure. Still, some reduction in variance was deemed worthwhile if it could be achieved through limited changes to individual weight values. A weight-trimming procedure was thus followed that changed weight values one at a time to achieve the largest reduction in variance and stopping the procedure with as few values as possible being changed.

Individual weight values and their components (that is, $w_{1i}$, $w_{2i}$, $w_{3i}$, $w_{4i}$) were listed for these cells and components for preliminary final weight values greater than two times the mean weight value for the cell inspected. The nonresponse adjusted base weight for selected cases was trimmed to the next largest nonresponse adjusted base weight value and the poststratification weights recomputed following trimming. The potential increase in variance was

recomputed. Weighted estimates for eight variables monitored during data collection were also computed and compared with the weighted estimates obtained with the preliminary final weights. The relative change in 1+*L* was compared with the relative change in the weighted estimates for the cell where trimming occurred. As long as the change in 1+*L* was greater than the relative change in the square of the weighted estimate, trimming was accepted as satisfactory. Trimming stopped in one cell when the relative change in weighted estimates became too large compared with the change in 1+*L*. A total of 49 weight values were trimmed in two rounds of trimming following this process.

## Final Weight Summary

Table T summarizes the distribution of the final trimmed weights for each of the 18 subclasses of interest. The female other race 25–44 year cell had the largest potential increase in variance due to weighting. Given the large sample size in the cell, and that substantial additional trimming would have been required to reduce the 1+*L* factor further, this somewhat larger 1+*L* factor was accepted.

There is large variation in final weights for the NSFG Cycle 6 shown in table T, and several extremely large weights. The variation and large weight values may be due to any one of nine weighting components in the final weight: six adjustments for unequal probabilities in the base weight, two adjustments for nonresponse propensity, and a final adjustment for poststratification. The largest contributing factor to the variation in and largest final weights is the adjustment for within-household selection probabilities and the substantial oversampling of 15–19-year-olds. Nonresponse adjustment and poststratification contribute little to the final weight variation.

Some of the variability has been trimmed, reducing the variation substantially. However, trimming was only done if the reduction of a large weight value substantially decreased the potential increase in variance (that

**Table Q. Nine logistic regression models predicting response to the main interview (Phases 1–3): National Survey of Family Growth, Cycle 6**

| Parameter | Estimate | Standard error | Pr > ChiSq |
|---|---|---|---|
| **Male** | | | |
| 15–19 years (n=1,330): | | | |
| Intercept | –12.0367 | 626.1000 | 0.9847 |
| Everresist | –3.5183 | 0.2451 | <.0001 |
| Everquest | 0.7093 | 0.2100 | 0.0007 |
| Everdelay | –0.7803 | 0.2177 | 0.0003 |
| Evercontact | 15.3092 | 626.1000 | 0.9805 |
| SCR_LANG | –0.8422 | 0.2943 | 0.0042 |
| SCR_SINGLEHH | 0.4381 | 0.8270 | 0.5963 |
| 20–24 years (n =1,169): | | | |
| Intercept | –12.3111 | 520.3000 | 0.9811 |
| Everresist | –3.0400 | 0.2502 | <.0001 |
| Everquest | 0.8035 | 0.1864 | <.0001 |
| Everdelay | –1.1714 | 0.1940 | <.0001 |
| Evercontact | 15.6769 | 520.3000 | 0.9760 |
| SCR_LANG | –1.2964 | 0.2316 | <.0001 |
| SCR_SINGLEHH | –0.1679 | 0.2332 | 0.4716 |
| 25–44 years: | | | |
| Black and Hispanic (n =1,594): | | | |
| Intercept | –13.3214 | 577.0000 | 0.9816 |
| Everresist | –2.4920 | 0.1994 | <.0001 |
| Everquest | 1.1283 | 0.1593 | <.0001 |
| Everdelay | –1.5796 | 0.1595 | <.0001 |
| Evercontact | 15.9141 | 577.0000 | 0.9780 |
| SCR_LANG | –0.7933 | 0.1530 | <.0001 |
| SCR_SINGLEHH | 0.3105 | 0.1841 | 0.0917 |
| Other (n=2,224): | | | |
| Intercept | –1.5086 | 1.0967 | 0.3344 |
| Everresist | –3.6956 | 0.1670 | <.0001 |
| Everquest | 1.0356 | 0.1414 | <.0001 |
| Everdelay | –1.0319 | 0.1422 | <.0001 |
| Evercontact | 4.8097 | 1.0316 | <.0001 |
| SCR_LANG | –1.8972 | 0.3786 | <.0001 |
| SCR_SINGLEHH | 0.0280 | 0.1481 | 0.8499 |
| **Female** | | | |
| 15–19 years (n =1,391): | | | |
| Intercept | –12.4894 | 767.8000 | 0.9870 |
| Everresist | –3.6254 | 0.2364 | <.0001 |
| Everquest | 0.5101 | 0.1997 | 0.0107 |
| Everdelay | –1.0298 | 0.2089 | <.0001 |
| Evercontact | 15.8693 | 767.8000 | 0.9835 |
| SCR_LANG | –0.7978 | 0.3124 | 0.0106 |
| SCR_SINGLEHH | –0.6299 | 0.5844 | 0.2811 |
| 20–24 years (n =1,672): | | | |
| Intercept | –14.1156 | 584.7000 | 0.9807 |
| Everresist | –3.3594 | 0.2282 | <.0001 |
| Everquest | 0.9440 | 0.1773 | <.0001 |
| Everdelay | –1.2181 | 0.1868 | <.0001 |
| Evercontact | 17.2140 | 584.7000 | 0.9765 |
| SCR_LANG | –0.9392 | 0.2155 | <.0001 |
| SCR_SINGLEHH | 0.2310 | 0.2850 | 0.4176 |
| 25–44 years: | | | |
| Black (n =1,291): | | | |
| Intercept | –12.2802 | 644.9000 | 0.9848 |
| Everresist | –3.4732 | 0.2449 | <.0001 |
| Everquest | 0.6225 | 0.2071 | 0.0027 |
| Everdelay | –1.0455 | 0.2094 | <.0001 |
| Evercontact | 16.7438 | 644.9000 | 0.9793 |
| SCR_LANG | –2.0470 | 1.0116 | 0.0430 |
| SCR_SINGLEHH | 0.0104 | 0.2719 | 0.9694 |

See footnotes at end of table.

**Table Q. Nine logistic regression models predicting response to the main interview (Phases 1–3): National Survey of Family Growth, Cycle 6—Con.**

| Parameter | Estimate | Standard error | Pr > ChiSq |
|---|---|---|---|
| Female—Continued | | | |
| 25–44 years: | | | |
| Hispanic (n =1,282): | | | |
| Intercept | –13.1382 | 575.7000 | 0.9818 |
| Everresist | –2.7555 | 0.2220 | <.0001 |
| Everquest | 0.5946 | 0.1736 | 0.0006 |
| Everdelay | –1.2686 | 0.1806 | <.0001 |
| Evercontact | 16.5567 | 575.7000 | 0.9771 |
| SCR_LANG | –0.9882 | 0.1613 | <.0001 |
| SCR_SINGLEHH | –0.2504 | 0.3387 | 0.4597 |
| Other (n=3,862): | | | |
| Intercept | –0.8169 | 1.0619 | 0.4417 |
| Everresist | –3.6215 | 0.1329 | <.0001 |
| Everquest | 0.8117 | 0.1152 | <.0001 |
| Everdelay | –1.1168 | 0.1157 | <.0001 |
| Evercontact | 4.6207 | 1.0507 | <.0001 |
| SCR_LANG | –1.5114 | 0.1739 | <.0001 |
| SCR_SINGLEHH | –0.2867 | 0.1568 | 0.0675 |

[1]Everresist: during any contact with the household, any respondent refused to participate in the survey.
[2]Everquest: during contact with the household, any respondent asked a question about the content or objectives of the survey.
[3]Everdelay: during any contact with the households, any respondent made an appointment for interview at another time.
[4]Evercontact: during any contact with the household, any respondent was contacted and provided information about the household.
[5]SCR_LANG: language of interview as indicated on screener interview.
[6]SCR_SINGLEHH: single eligible person in household, recorded on screener interview.

is, a substantial reduction in the 1+$L$ factor). Large weight values remain but they do not contribute as substantially to the potential increase as weight values that have already been trimmed.

**Table R. Mean, minimum, and maximum combined nonresponse adjustments, and potential increase in variance due to weighting, by age group, sex, and race or ethnicity: National Survey of Family Growth Cycle**

| Sex, race or ethnicity, and age group | Mean adjustment | Minimum adjustment | Maximum adjustment | Increase in variance (1+$L$) |
|---|---|---|---|---|
| Female | | | | |
| Black: | | | | |
| 15–19 | 1.2005 | 1.0517 | 2.5520 | 1.0368 |
| 20–24 | 1.2592 | 1.0652 | 2.0933 | 1.0340 |
| 25–44 | 1.3135 | 1.0519 | 2.5520 | 1.0488 |
| Hispanic: | | | | |
| 15–19 | 1.2334 | 1.0485 | 2.5520 | 1.0369 |
| 20–24 | 1.3410 | 1.0676 | 2.5520 | 1.0462 |
| 25–44 | 1.3593 | 1.0570 | 2.5520 | 1.0485 |
| Other: | | | | |
| 15–19 | 1.2066 | 1.0514 | 2.5520 | 1.0315 |
| 20–24 | 1.2767 | 1.0556 | 2.5520 | 1.0394 |
| 25–44 | 1.2908 | 1.0662 | 2.5520 | 1.0414 |
| Male | | | | |
| Black: | | | | |
| 15–19 | 1.2235 | 1.0614 | 2.1015 | 1.0347 |
| 20–24 | 1.2412 | 1.0546 | 2.5520 | 1.0327 |
| 25–44 | 1.2553 | 1.0555 | 2.5520 | 1.0425 |
| Hispanic: | | | | |
| 15–19 | 1.2564 | 1.0560 | 2.1015 | 1.0384 |
| 20–24 | 1.3033 | 1.0519 | 2.5520 | 1.0386 |
| 25–44 | 1.3458 | 1.0574 | 2.5520 | 1.0512 |
| Other: | | | | |
| 15–19 | 1.2152 | 1.0541 | 2.5520 | 1.0335 |
| 20–24 | 1.2465 | 1.0543 | 2.5520 | 1.0361 |
| 25–44 | 1.2526 | 1.0527 | 2.5520 | 1.0394 |

# Estimating Equation for Totals

The NSFG Cycle 6 estimator of the number of men or women with a particular characteristic can be computed as

$$\hat{N} = \sum_i w_{5i} I_i$$

where $w_{5i}$ is the final trimmed weight for the *ith* sample person and $I_i = 1$ if the person has the characteristic and $I_i = 0$ otherwise.

To estimate the total number of events or the total for a particular variable, such as the total number of live births, for persons in a particular subgroup identified by the indicator $I_i$ is computed as

$$\hat{Y} = \sum_i w_{5i} I_i y_i$$

where $y_i$ is the value of the characteristic of interest. The variable $y_i$ may be a count such as number of live births or a continuous measure such as income.

Estimates of means or proportions can be constructed from these estimated totals. For example, the proportion of the population with a characteristic of interest can be computed as the ratio estimator

$$p = \frac{\sum_i w_{5i} I_i}{\sum_i w_{5i}}$$

Alternatively, the mean for a variable $y_i$ can be computed as the ratio estimator

$$\bar{y} = \frac{\sum_i w_{5i} I_i y_i}{\sum_i w_{5i} I_i}$$

Sampling variances for these estimators must be computed taking the stratified multistage sample design into account. As mentioned before, standard statistical software does not properly account for these design features in estimation. Variance estimation procedures are described in a later section and use of software that accounts for weighting, stratification, and clustering in estimation of variances is illustrated.

# Imputation Procedures

## Background on Imputation for Missing Values

In every survey some values are missing for some variables. These missing values create inconsistencies in reported results, since an analysis using different variables with different numbers of missing values will typically have unequal sample sizes. Item missing values may also contribute to bias if the analysis is restricted to complete cases, those with valid values. If missing values differ on average from those observed in the valid cases, an analysis that only uses cases with complete data may yield biased estimates.

There are several remedies for item missing values, including weighting specific to each analysis and replacement of missing values with predicted values. The latter, imputation, will make analyses more consistent and may reduce bias due to item nonresponse (depending on the predictive power of the imputation procedure).

However, in a survey as complex as NSFG Cycle 6 with thousands of variables, imputation may be required for a large number of variables. When, as is typically the case, only a few values are missing for the overwhelming majority of variables, imputation can be an expensive process. Each variable to be imputed requires detailed specification of how predicted values to be assigned to replace missing values will be estimated or obtained.

As in previous NSFG cycles, imputation was used only for a subset of variables, those to be used in the primary tabulations and analyses to be released by NCHS. There were three files for which variables required imputation: female, male, and pregnancy. Each had somewhat different sets of variables. An extensive set of recoded variables was chosen for imputation. These recoded variables were usually combinations of two or more questions from the questionnaires, and they were judged to be necessary for the production of NCHS reports from the NSFG. A list of these recoded variables is provided in the documentation. There were a total of 356 variables identified for imputation across the three data files: 253 in the female file, 78 in the male file, and 25 in the pregnancy file. These variables and the type and number of cases imputed are shown in the tables in Appendix II of this report. For all variables with imputed values, a set of indicator variables is available in the female, male, and pregnancy files that designate which cases received imputation for each variable. These "imputation flags" may be used by the analyst to drop imputed values from an analysis or to replace the imputed values in the data set with values predicted from a model or procedure specified by the user.

There is a considerable range of methods used to impute for item

**Table S. Poststratification adjustment factors for 36 groups by age, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6**

| Sex, race or ethnicity, and age group | Census population plus DMDC population | Sum of NSFG Cycle 6 nonresponse adjusted weights | Post-stratification adjustment factor |
|---|---|---|---|
| Female | 61,560,728 | 40,393,425 | . . . |
| Black | 8,586,724 | 6,287,567 | . . . |
| 15–19 | 1,498,790 | 1,125,827 | 1.3313 |
| 20–24 | 1,457,657 | 960,712 | 1.5173 |
| 25–29 | 1,302,135 | 965,773 | 1.3483 |
| 30–34 | 1,392,656 | 1,071,172 | 1.3001 |
| 35–39 | 1,455,780 | 951,197 | 1.5305 |
| 40–44 | 1,479,706 | 1,212,886 | 1.2200 |
| Hispanic | 9,106,748 | 7,870,993 | . . . |
| 15–19 | 1,520,555 | 1,301,782 | 1.1681 |
| 20–24 | 1,632,281 | 1,453,466 | 1.1230 |
| 25–29 | 1,654,328 | 1,525,397 | 1.0845 |
| 30–34 | 1,594,616 | 1,256,541 | 1.2691 |
| 35–39 | 1,447,511 | 1,338,191 | 1.0817 |
| 40–44 | 1,257,457 | 995,616 | 1.2630 |
| Other race | 43,867,256 | 26,234,865 | . . . |
| 15–19 | 6,814,699 | 3,874,401 | 1.7589 |
| 20–24 | 6,749,458 | 4,798,243 | 1.4067 |
| 25–29 | 6,293,017 | 3,741,572 | 1.6819 |
| 30–34 | 7,285,198 | 4,198,948 | 1.7350 |
| 35–39 | 7,949,871 | 4,356,281 | 1.8249 |
| 40–44 | 8,775,013 | 5,265,420 | 1.6665 |
| Male | 61,146,887 | 35,440,295 | . . . |
| Black | 7,277,915 | 5,178,239 | . . . |
| 15–19 | 1,473,555 | 1,081,183 | 1.3629 |
| 20–24 | 1,250,646 | 785,146 | 1.5929 |
| 25–29 | 1,047,040 | 915,933 | 1.1431 |
| 30–34 | 1,122,586 | 859,530 | 1.3060 |
| 35–39 | 1,171,917 | 841,770 | 1.3922 |
| 40–44 | 1,212,171 | 694,677 | 1.7449 |
| Hispanic | 10,188,079 | 7,614,180 | . . . |
| 15–19 | 1,627,708 | 1,426,891 | 1.1407 |
| 20–24 | 1,951,833 | 1,380,156 | 1.4142 |
| 25–29 | 1,935,434 | 1,485,032 | 1.3033 |
| 30–34 | 1,789,473 | 1,147,133 | 1.5600 |
| 35–39 | 1,571,123 | 1,185,771 | 1.3250 |
| 40–44 | 1,312,508 | 989,197 | 1.3268 |
| Other race | 43,680,893 | 22,647,876 | . . . |
| 15–19 | 7,106,686 | 3,933,169 | 1.8069 |
| 20–24 | 6,680,687 | 4,349,801 | 1.5359 |
| 25–29 | 6,243,607 | 3,298,239 | 1.8930 |
| 30–34 | 7,225,525 | 3,734,755 | 1.9347 |
| 35–39 | 7,813,845 | 3,662,293 | 2.1336 |
| 40–44 | 8,610,543 | 3,669,619 | 2.3464 |

. . . Category not applicable.

NOTE: DMDC is Defense Manpower Data Center and NSFG is National Survey of Family Growth.

**Table T. Mean, minimum, and maximum final (trimmed) weight, and potential increase in variance due to weighting, by age group, sex, and race or ethnicity: National Survey of Family Growth, Cycle 6**

| Sex, race or ethnicity, and age | Mean weight | Minimum weight | Maximum weight | Increase in variance (1+*L*) |
|---|---|---|---|---|
| Female | | | | |
| Black: | | | | |
| 15–19 years | 6,193.45 | 1,143.98 | 25,581.37 | 1.4089 |
| 20–24 years | 6,686.64 | 1,238.52 | 63,141.59 | 1.8866 |
| 25–44 years | 5,261.97 | 579.5 | 33,290.92 | 1.5971 |
| Hispanic: | | | | |
| 15–19 years | 6,582.73 | 1,460.02 | 40,128.20 | 1.4732 |
| 20–24 years | 5,570.79 | 1,407.22 | 28,985.12 | 1.3738 |
| 25–44 years | 5,590.57 | 610.2 | 41,698.26 | 1.5477 |
| Other: | | | | |
| 15–19 years | 10,066.00 | 411.45 | 49,916.15 | 1.3274 |
| 20–24 years | 7,922.17 | 727.85 | 101,214.07 | 1.5886 |
| 25–44 years | 10,117.77 | 118.66 | 261,879.95 | 1.8732 |
| Male | | | | |
| Black: | | | | |
| 15–19 years | 7,188.02 | 1,690.83 | 26,374.42 | 1.5231 |
| 20–24 years | 8,625.23 | 733.93 | 34,284.43 | 1.7284 |
| 25–44 years | 7,851.04 | 494.85 | 48,561.20 | 1.9309 |
| Hispanic: | | | | |
| 15–19 years | 6,926.19 | 1,413.10 | 29,573.44 | 1.4179 |
| 20–24 years | 9,078.21 | 662.56 | 59,698.43 | 1.6788 |
| 25–44 years | 9,819.59 | 1,643.42 | 61,190.55 | 1.8014 |
| Other: | | | | |
| 15–19 years | 10,435.89 | 2,251.27 | 67,365.59 | 1.4039 |
| 20–24 years | 11,558.58 | 1,878.62 | 68,059.86 | 1.5941 |
| 25–44 years | 18,498.43 | 746.43 | 109,064.17 | 2.0273 |

missing values in survey practice (13). In some instances, values can be replaced through a logical derivation based on the relationship among variables. In most cases, imputation must be based on a form of prediction. The imputation procedures used in survey practice range from simple cell-mean procedures to sequential hot deck, flexible matching, and regression imputation (13).

*Imputation in NSFG Cycle 5*—Previous NSFG cycles have used a combination of logical, hot deck, and regression imputation procedures. For example, most imputation in Cycle 5 used a sequential hot deck procedure (5). The Cycle 5 hot deck imputations were based on a single-variable sequential procedure for each variable. For each variable to be imputed, data were sorted by one or more "predictor" variables (many sort variables were used for more than one variable). The sort variables were chosen from among those observed to be most highly correlated with the variable to be imputed.

The data were then grouped into imputation classes (sometimes referred to as matrices), and the sequential imputation begun. In the sort order, if a case had a nonmissing value for the variable being imputed, the value was stored as the "hot" value. If a case was missing for the variable being imputed, the "hot" value was substituted, or imputed, for the missing value.

Variables were imputed in a selected order to allow some variables to be used as sorting or classing variables which themselves had missing values. Thus, the sorting and classing variables had to be imputed before being used in the imputation of another variable.

The hot deck imputation procedure used in Cycles 3–5 imputed one variable at a time and used only a limited set of sort or class variables as predictors. In addition, hot deck imputation was completed for some variables before others, and the imputed values for some variables were used in sorting or classifying other subsequently imputed variables. Thus, there was an order of imputation imposed that could limit the range of possible values that were imputed for some variables. It is now known, however, that these features may have an undesirable effect on the results obtained from imputed data. For example, single-variable (as opposed to multivariable) imputations may disrupt or attenuate an association structure, reducing correlations among variables in the imputed data.

*Imputation in Cycle 6*—In Cycle 6, most imputation was performed using a regression procedure. The procedure was programmed to impute all variables within a data file simultaneously, using a large number of potential predictors for each imputed variable. The procedure employed a generalized linear regression model, providing for the imputation of interval, dichotomous, ordinal, nominal, and count scales for the variable being predicted.

A small methodological investigation compared Cycle 5 sequential hot deck imputed values for a selected set of variables to sequential regression imputations for those variables. Findings indicated that the sequential regression imputation procedure generated imputed values with similar distributions to those obtained by the hot deck and did so while preserving covariance structure. The sequential regression procedure is available in publicly available software from the ISR Web site at the University of Michigan Web site: http://www.isr.umich.edu/src/smp/ive/.

IVEware was used to generate most of the imputed values for Cycle 6. Logical imputation was also used, where appropriate, sometimes for the same variables receiving sequential regression imputations.

# Sequential Regression Imputation

Regression imputation was implemented using a sequential regression method described in Raghunathan et al. (14). The method employs both a sequential procedure to impute multiple variables in a particular order and Bayesian methods to perturb regression coefficients used in predicting values and to add residuals to imputed values. The method allows for

alternative scales for each variable to be imputed, including linear, dichotomous, polytomous, and count.

In the latter three scales, predicted values can be obtained that are outside the range of acceptable values. Under logistic regression, predicted proportions between zero and one are generated. These values are converted to zero or one values through a random process. For example, suppose that the predicted value for a dichotomous variable is 0.4. A uniform random number from zero to one is generated. If the random number is less than 0.4, the imputed value of "1" is assigned. Otherwise, the imputed value of "0" is assigned. Similar random assignment of valid values occurs for polytomous and count variables.

Imputed values from other variables are used in the prediction of values used as imputations for a given variable. The procedure is multivariate, using all variables specified as predictors of each variable in the imputation procedure and thus effectively preserves existing covariance structure in the data.

Consider a set of $p+q$ variables from NSFG Cycle 6, where the $p$ variables do not require imputation (all values were observed, or logically imputed prior to regression imputation). The $q$ variables may be ordered from largest to smallest number of cases requiring imputation, although this ordering is not necessary and was not employed in Cycle 6 imputation.

The imputation requires several cycles. In the first, the first of the $q$ variables, $y_{p+1}$, is regressed on all the $p$ variables without missing values. A linear, logistic, multinomial, or Poisson regression is used, depending on the measurement scale for $y_{p+1}$ (the variable being imputed) specified by the user. From this regression, a predicted value $\hat{y}_{p+1}$ is generated for each missing value. Still in the first cycle, the second variable, $y_{p+2}$, is regressed on the $p$ variables without missing values and $y_{p+1}$ (including its imputed values). Predicted values for $y_{p+2}$ from this estimated regression model are assigned to item missing values. This cycle continues by repeating the regression imputation process for the variables $y_{p+3}$, ..., $y_{p+q}$, imputing in each case using all $p$ variables with no missing values and all imputed variables obtained to that point.

Cycle 2, 3, and so on (a total of five cycles are typically used) repeat the regression imputation, but include all variables, the $p$ with no missing values and the other $q$–1 with imputed values, as predictors. Thus, $y_{p+1}$ is regressed on the set of predictors $y_1$, ..., $y_p$, $y_{p+2}$, ..., $y_{p+q}$, and imputed values obtained from the predicted values under the regression model.

The predicted values from each regression in each cycle are obtained through a process that perturbs the coefficients and adds a random residual. The process is described in the appendix of Raghunathan et al. (14).

Variables may have restrictions to limit the set of cases to be imputed. For example, persons who have never had sex should not receive an imputed value for age at first intercourse. Restrictions are readily programmed in the software. Values to be imputed may be bounded within specified limits. Bounds for imputed values can also be set, and truncated distributions are used to assign predicted values. In the NSFG Cycle 6, a major part of the work of imputation was in specifying appropriate bounds for imputed values because of the extensive interrelations in the data.

The sequential regression imputation method is implemented in the IVEware software system (http://www.isr.umich.edu/src/smp/ive/). Only a single run of this software was used to impute missing values for all recode variables in each data set simultaneously.

# Multiple Imputation Using Sequential Regression

The IVEware system also can perform multiple imputation, a method that allows users to obtain valid estimates of precision, test statistics, and statistical significance for imputed data. This feature was not used with the NSFG Cycle 6 imputations, but may be used by NSFG users in Cycle 6 or in the future, so a brief description of multiple imputation is included.

Instead of a single imputation, the imputation process is repeated, say $m$=5 times. The multiple imputation thus yields not one but $m$=5 data sets with valid values repeated and imputed values that differ among data sets. Variation among data sets can be combined with variation within data sets to account for the effect of imputation on variances.

Suppose, for example, that $\hat{\beta}^{(l)}$ represents a vector of estimated regression coefficients obtained from one of the five imputed data sets, say from the $l$-th multiple imputation. Let $V^{(l)}$ denote the covariance matrix for the coefficients for the $l$-th imputed data set.

The multiply imputed estimate of the regression coefficients is obtained as

$$\hat{\beta} = \sum_{l=1}^{m} \hat{\beta}^{(l)} / m,$$

the average of the $m$ estimated regression model coefficients. The covariance of the regression coefficients is obtained as

$$\hat{V} = \left(\sum_{l=1}^{m} \hat{V}^{(l)} \Big/ m\right) + \left(\frac{m+1}{m}\right) \hat{\beta}, \text{ where}$$

$$\hat{\beta} = \sum_{l=1}^{m} (\hat{\beta}^{(l)} - \hat{\beta})\,(\hat{\beta}^{(l)} - \hat{\beta})^t \Big/ (m-1).$$

These estimation procedures are implemented in a series of analytic procedures (including linear and logistic regression) within IVEware.

The NSFG Cycle 6 data were obtained from a weighted stratified multistage sample of eligible persons. It is most appropriate if the estimates $V^{(l)}$ are obtained from a procedure that accounts properly for these factors. IVEware implements multiple imputation estimation using an estimate $V^{(l)}$ obtained from a jackknife replication procedure employing NSFG strata and PSU variables and weighted estimates.

## Regression Imputation in NSFG Cycle 6

Tables I (female), II (male), and III (pregnancy) in Appendix II show the variables imputed in Cycle 6, the number of regression imputed values, and the number of logically imputed values. Extended names and definitions for each variable are shown in the public-use documentation.

In the female file, there are 7,643 records. The largest number of imputed values for any variable occurred for POVERTY (family income as a percent of the poverty level) and TOTINCR (total family income). Both received 630 imputed values, 7.6 percent of all values for these variables. All of these were from the sequential regression imputation. Other variables had higher percentages imputed but these were typically for variables that had very few occurrences in the data set. For example, AGEPREG13, the age at the 13$^{th}$ mentioned pregnancy, has four values, one of which is imputed. In most variables, the amount of imputation is only a few cases.

Other variables had logical as well as regression imputed values assigned. For example, AGEBABY1 (the woman's age when her first baby was born) in the female file had 31 logical and 20 regression imputed values. The logical imputations were assigned on the basis of the date of birth of the mother and the child. Regression imputations were assigned when those dates or other information needed to logically assign a value were missing.

In the male and pregnancy files, similar results are observed. For example, among the 4,928 male records, POVERTY and TOTINCR have the largest number of imputed values, 413, or 8 percent. In the pregnancy file, FMAROUT5 (formal marital status at pregnancy outcome) had the largest number of imputed values (all from regression imputation), with 8.2 percent of the 13,240 values imputed.

# Variance Estimation

## Background

Since NSFG Cycle 6 estimates are based on a sample rather than a complete enumeration of the eligible population, they are subject to error, a departure between the true population value and the estimated value. This difference may be due to systematic or fixed sources of error, such as nonresponse or noncoverage bias. It is also attributable to variable sources of error, including the use of a sample to represent the population.

Probability sampling, as used in the NSFG Cycle 6, allows the direct estimation of one of these sources of error, the variable error due to sampling. A considerable share of the survey design and estimation literature develops proper procedures for estimating the sampling variance under different sample selection techniques. See Wolter (15) for a review of variance estimation techniques.

In this section, the focus is on the estimation of sampling variance for NSFG Cycle 6 estimates that properly account for the principal effects of the different sampling techniques employed in the NSFG. There are four principal design features that can be accounted for in Cycle 6: stratification of PSUs, selection of PSUs (cluster sample selection), weights which adjust for unequal probabilities of selection and other design features, and the presence of imputed values.

The proper treatment of imputed values in variance estimation can be done through multiple imputation. However, because of the size of the data files, multiply imputed values have not been released for the NSFG Cycle 6. A future publication may explore the extent to which sampling variances are altered by imputed values. Such an investigation is beyond the scope of this report.

The remainder of this section describes how variance estimates may be obtained from Cycle 6 that properly account for the remaining three features: stratification, clustering, and weighting. The discussion considers the effect of these features on estimated sampling variances, the availability of software that can estimate variances taking these features into account, the creation of codes that can be used for variance estimation, and illustrations of computation of sampling variances from the Cycle 6 data using four different software systems.

In these illustrations, the magnitude of the effects of the complex design, the stratification, clustering, and weighting, are presented. Usually, the sampling variances computed taking the design into account are larger than those obtained using methods that do not account for these features. This increase is the reason that analysts must consider carefully how to compute estimates of variance from the NSFG.

If software is used that does not account for the complex design, sampling variances will often be underestimated. The consequence will be test statistics that generate smaller probability test values than is actually correct and confidence intervals that are narrower than they should be. Both of these types of errors can lead to incorrect inferences drawn from the data. For example, smaller probability test levels might lead an analyst to reject a null hypothesis incorrectly.

For this reason, all users of NSFG Cycle 6 data should use variance estimation procedures discussed in this section. They will assure that incorrect test procedures or overly narrow confidence intervals are not used in drawing conclusions from the Cycle 6 data.

## Summary of Variance Estimation Principles

There are two primary approaches used to estimate sampling variance from complex sample surveys: Taylor series approximation methods and repeated replication methods. The former are based on an analytic treatment of statistical estimates, while the latter are based on computer intensive resampling of the survey data.

Complex sample survey data generate statistics which, through the use of weights, can be estimated by standard statistical software. That is, the estimated percentages, means, regression coefficients, and other statistics are computed properly using weighted data. However, for the purpose of estimating test statistics and standard errors of estimates, standard software treats the data as though the sample was selected using very simple sampling methods. The consequence, as noted previously, is an underestimation of standard errors and an overstatement of the significance of test statistics.

The Taylor series approximation, or linearization procedure, can be adapted to account for the complex design features. Taylor series linearization is a widely used technique in statistics for obtaining approximate variance estimates for nonlinear statistics. The Taylor series is an infinite series expansion of a function that has a valuable property: for ''well-behaved'' functions, the successive terms in the expansion rapidly approach zero. Thus, for some functions, the first term of the Taylor series expansion can be used to represent the entire infinite series.

The first term is linear in the variables in the function. Variances are computed for the ''linearized'' function. The properties of the Taylor series expansion approximation variance estimates are well understood empirically (16). It has been applied to a wide range of statistics computed from sample survey data; from means and proportions to logistic regression coefficients estimated using iterative numerical solutions.

For the NSFG Cycle 6 data, the Taylor series method can be adapted to the paired sampling error computing units (SECUs) created to represent the stratification and clustering in the sample. Paired differences variance estimation procedures are widely implemented in software systems that implement Taylor series approximations for complex sample survey designs.

The repeated replication procedures are based on early work of Mahalanobis (17) and Deming (18). They proposed and encouraged the use of interpenetrating or replicating sampling methods, which simplified variance estimation considerably. For example, if a sample were comprised of $T$ replicate samples, each selected in the same way, and a statistic $\hat{\theta}_t$ was computed from each replicate, the sampling variance of the average

$$\bar{\theta} = \frac{1}{T}\sum_{t=1}^{T}\hat{\theta}_t$$

can be computed simply as

$$var\,(\bar{\theta}) = \frac{1}{T(T-1)}\sum_{t=1}^{T}(\hat{\theta}_t - \bar{\theta})^2.$$

The idea of replicated sampling was extended to survey data by McCarthy (19), who proposed selecting replicate samples from a larger sample by employing in the replicate sampling procedures, selection methods that were identical to those used in the original sample selection. For example, in the NSFG Cycle 6, replicate half-samples would be drawn by selecting a single SECU from each stratum, and taking all cases within a SECU, to comprise the replicate sample. Alternatively, a ''jackknife'' replication procedure ''drops one'' SECU in a stratum, and retains all other SECUs and their cases in the sample. In the NSFG Cycle 6, the cases in the SECU remaining in the stratum from which the one SECU was dropped must have their values increased (doubled) to account for the underrepresentation of the stratum.

With half-sample repeated replication, the sampling variance for a statistic can be estimated in a form very similar to that given previously for replicated sampling. Let $\tilde{\theta}$ denote the estimate of the statistic of interest computed from the entire sample. Then

$$var(\tilde{\theta}) = \frac{1}{T}\sum_{t=1}^{T}(\hat{\theta}_t - \tilde{\theta})^2,$$

where $\hat{\theta}_t$ is computed from the $t$-th replicate. For the jackknife repeated replication procedure,

$$var(\tilde{\theta}) = \sum_{t=1}^{T}(\hat{\theta}_t - \tilde{\theta})^2.$$

See Rust (20) for a more detailed review of these procedures.

## Variance Estimation Software

These variance estimation procedures are all implemented in a number of computer software packages that are commercially available, or available for free download over the Internet. In earlier NSFG cycles, it was often complex and difficult to obtain and use variance estimation software, but now there are several software packages that allow users to estimate variances for means, proportions, regression coefficients, logistic regression coefficients, and other statistics. These statistical software packages implement the Taylor series approximation, repeated replication methods, or both approaches for a complex sample survey like the NSFG Cycle 6.

There are two basic types of software systems available for this purpose: stand-alone and integrated packages. The stand-alone software packages require users to input data into a special format used by the system. For example, WesVar is a stand-alone package that implements repeated replication procedures and requires data be converted to the WesVar format for estimation. Fortunately, WesVar is an easy to use window-driven system that can read data directly from a number of standard statistical software systems, such as SPSS. Other stand-alone systems require more time-consuming conversion of data to their particular format and are based on command input rather than a ''point and click menu'' system.

Integrated software for estimation from complex sample survey data allows a user to conduct an analysis without converting data to another format. The user can construct a survey data set in a format used by a major statistical software package, such as SAS or STATA, and then compute estimates and test statistics within those systems that take complex design features into account. For example, SAS version 9 has, at the time of this writing, four procedures that properly account for complex features such as stratification, clustering, and weighting.

STATA has a larger set of survey or "svy" versions of frequently used analysis commands that account for stratification, clustering, and weighting in complex sample surveys. SUDAAN is a stand-alone system that comes in a SAS compatible version, which allows SUDAAN commands to be embedded in SAS programs as though they were part of the SAS language. SUDAAN implements a wider range of statistical procedures for complex sample survey data than are currently available in SAS version 9. Users with SAS data sets can thus analyze data using either SAS or SUDAAN, although they will find more analysis procedures available in SUDAAN. Finally, SPSS has recently announced the introduction of variance estimation procedures as an optional module. As of this writing, the SPSS optional module has not been tested on NSFG Cycle 6 data, nor have examples been developed for users to illustrate implementation on Cycle 6 data.

NSFG Cycle 6 users can find descriptions of these and other software systems for estimation from complex sample survey data, along with detailed explanations of their features, at a Web site maintained by the Survey Research Methods Section of the American Statistical Association, currently located at http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html. All NSFG Cycle 6 users are strongly encouraged to visit this Web site and become more familiar with the features of these software systems.

Empirical comparisons of these software packages have shown that they provide identical estimates of statistics of interest, and that the estimated sampling variances of the estimated statistics are virtually the same, regardless of the computational method used to estimate the sampling variances. That is, all four systems yield nearly identical estimates.

The choice among these systems can thus be based on practical considerations of convenience, cost of software acquisition and maintenance, and current data format.

For example, a user with data already in SAS format, or working in an environment where SAS is widely used, would find SAS and SUDAAN to be most convenient to use. SUDAAN can compute variances for some types of analyses that SAS cannot, but SUDAAN must be purchased and the license must be renewed each year. A free software system that operates within SAS is also available, IVEware. It can be downloaded from a Web site at the University of Michigan, together with documentation for installation and use: www.isr.umich.edu/src/smp/ive/. IVEware has as wide a range of variance estimation procedures as SUDAAN.

Users with data in SPSS format will find WesVar a convenient package to use since WesVar converts SPSS data into its own internal format through a few simple windows. Alternatively, SPSS data may be converted to SAS, STATA, or another format through any number of file conversion systems, and then used by another complex sample survey package such as SUDAAN.

Users with data in STATA format can apply the many "svy" versions of STATA commands to the data directly. The STATA system does require specification of stratification, cluster, and weight variables through a design statement. The "svy" commands are completely integrated into the STATA command language.

## Application to the NSFG

### Sampling Error Computing Units

The NSFG sample has 121 PSUs. Each sample PSU was drawn from a stratum of one or more PSUs. Sampling variance cannot be estimated directly from this type of a design, since there is but one selection per stratum. The NSFG PSUs were arranged into a set of SECUs that could be grouped into pairs for variance estimation purposes.

The 28 self-representing (SR) PSUs were divided into two or four representative units each by a systematic sample of the segments within each PSU. For example, in a large SR PSU the segments were numbered in sample selection order with each sample domain as 1, 2, 3, 4, 1, 2, 3, 4, 1, 2 ... Segments with the same number were grouped together to form a pseudo-PSU representing the entire MSA. The first and third such "combined units" were paired to form a stratum with two combined units, or SECUs. Thus, the large PSU in this example has four SECUs in the data set, grouped into two pairs. Each pair is a pseudo-stratum. Smaller SR PSUs were divided into two SECUs, which are in turn grouped into a single pair as a pseudo-stratum. This SR PSU process generated a total of 38 stratum pairs of SECUs.

For the remaining 93 non-self-representing (NSR) PSUs (including the 11 unique Hispanic sample PSUs), the strata were inspected to identify pairs or other groupings that were as alike as possible. Two NSR PSUs were linked to form one pseudo-NSR PSU for purposes of variance estimation, yielding a total of 92 NSR pseudo PSUs. Criteria for deciding similarity included MSA status and geographic location. A total of 46 pairs of PSUs were created.

In combination with the SR PSU SECUs, there are thus a total of 84 pairs of SECUs or 168 SECUs. Two variables, SEST and SECU, are in the Cycle 6 data set to identify these groupings. The codes for these values have been randomly ordered to mask the identity of the units.

Appendix III contains tables showing the distribution of all 12,571 female and male cases and the 13,593 pregnancy records in the pregnancy file. By chance, one of the SECUs contained no cases in the pregnancy file: SEST 42, SECU 1. This anomaly would create difficulty for users seeking to compute sampling variances using some of the statistical software systems illustrated subsequently. Six of the 32 pregnancies in SECU 2 in SEST 42 were from one woman each. One of these was chosen at random and its SECU code for the pregnancy file changed to 1. Thus, there is now one pregnancy record in SEST 42, SECU 1. This one record has a different SECU code than originally assigned. Its presence will not cause any noticeable bias in estimated sampling variances but it will allow the computations to proceed.

## Sampling Variance Estimation Examples

To illustrate the convenience of using these software systems for the NSFG Cycle 6 data, nine example analyses have been developed and implemented on each of four software systems. These illustrations are shown in their entirety on the Internet at: www.cdc.gov/nchs/about/major/nsfg/nsfgvar.htm.

Selected examples are shown in Appendixes IV through VII of this report. The software systems used are SAS version 9.1 (which uses Taylor series approximation methods in several survey procedures), SUDAAN 8.0.2 (which uses both Taylor series approximation and repeated replication methods), and STATA version 8.0 (which uses Taylor series approximation methods). In the examples on the Internet, examples with WesVar version 4.1 (which uses repeated replication methods) are also shown.

The nine illustrations on the Internet site have been chosen to provide users with examples of how to compute sampling variances for NSFG Cycle 6 for contingency tables, tables of mean values, linear regression, logistic regression, and tables of estimated total counts. Some of the illustrations use a subset of data, such as an age group. Most of the variables used in the illustrations are selected from the extensive recoded variables available in NSFG Cycle 6. These recoded variables have no item missing values because the item missing values have been replaced with imputed values. Users can suppress the use of imputed values through the use of the imputation flags.

The illustration tables and analyses have at most three to five variables, reducing the complexity of recoding and software set up. The four examples shown in Appendixes IV, V, VI, and VII of this report contain the programs and output for:

1. Appendix IV: Female respondent file. Percentage of all women using the oral contraceptive pill, by 5 year age groups 15–19, 20–24, 25–29, 30–34, 35–39, and 40–44.

**SAS 9.1 Program**

```
data NSFG.EX1;
set  NSFG.FEMALES;
if 15 le AGER le 19 then agerx=1;
if 20 le AGER le 24 then agerx=2;
if 25 le AGER le 29 then agerx=3;
if 30 le AGER le 34 then agerx=4;
if 35 le AGER le 39 then agerx=5;
if AGER ge 40 then agerx=6;
if CONSTAT1=6 then pill=1; else pill=2;
run;

proc surveyfreq data=NSFG.EX1;
stratum SEST;
cluster SECU;
weight  FINALWGT;
table agerx*pill / row deff;
run;
```

**Figure 10. SAS 9.1 program code illustration 1**

2. Appendix V: Female respondent file. Mean number of children ever born (PARITY) per woman, for women 20–44, by race and Hispanic origin and age.
3. Appendix VI: Female respondent file. Multiple regression, regressing number of children ever born on race and ethnicity, age (single years), and education, for women 20–44.
4. Appendix VII: Male and female combined file. Logistic regression of the probability of strongly agreeing with the statement, "a young couple should not live together unless they are married" on age (single years), gender, and race and ethnicity, and education.

The NSFG Web site (http://www.cdc.gov/nchs/nsfg.htm) contains the following additional examples showing STATA, SAS, SUDAAN, and WesVar output:

1. Male file. Percentage of all men 20–44 years of age who have ever fathered one or more children, by race.
2. Male and female combined file. Percentage of all men and women 15–44 years of age, by gender and age (15–24, 25–34, 35–44) who strongly agree that "a young couple should not live together unless they are married."
3. Male and female combined file. Percentage of males and females 15–19 who have ever had sexual intercourse, by gender.
4. Pregnancy file. Percent distribution of births that were wanted, mistimed, or unwanted, by race and Hispanic origin and age at pregnancy outcome.
5. Female file. Estimates of the total number of women using the oral contraceptive pill, by 5-year age groups.

Programs and a subset of the output for the first illustration are presented here. The problem is to estimate the percent of women using the oral contraceptive pill, and to estimate that percentage for women in six 5-year age groups. The female data file is used for this analysis in all four packages, and in all four software systems a cross-tabulation procedure is used.

All four systems require some prior recoding of the variables to be used in the analysis. Since SAS version 9 and SUDAAN use the same SAS data set and recoding commands, the recodes are given only in the SAS program. While the SAS and STATA programs "use" the female data directly, the WesVar program has three windows that illustrate the conversion of an SPSS data file to WesVar format.

We begin with the SAS version 9 program and output from the SAS procedure SURVEYFREQ in figure 10. The program has a short recode in a DATA step to convert the continuous age variable AGER to a recoded variable with six age group categories and to code oral contraceptive use from CONSTAT1.

**SAS 9.1 Output**

The SURVEYFREQ Procedure

Data Summary

| | |
|---|---|
| Number of Strata | 84 |
| Number of Clusters | 168 |
| Number of Observations | 7643 |
| Sum of Weights | 61560714.8 |

Table of agerx by pill

| agerx | pill | Frequency | Weighted Frequency | Std Dev of Wgt Freq | Percent | Std Err of Percent | Design Effect | Row Percent | Std Err of Row Percent |
|---|---|---|---|---|---|---|---|---|---|
| 15-19 | Yes | 187 | 1633986 | 176138 | 2.6543 | 0.2740 | 2.2211 | 16.6155 | 1.4964 |
| | No | 963 | 8200123 | 308550 | 13.3204 | 0.4921 | 1.6029 | 83.3845 | 1.4964 |
| | Total | 1150 | 9834109 | 380244 | 15.9746 | 0.5744 | 1.8784 | 100.000 | |
| 20-24 | Yes | 424 | 3127289 | 338308 | 5.0800 | 0.4776 | 3.6146 | 31.7826 | 1.9966 |
| | No | 939 | 6712331 | 373170 | 10.9036 | 0.4710 | 1.7454 | 68.2174 | 1.9966 |
| | Total | 1363 | 9839620 | 621472 | 15.9836 | 0.7570 | 3.2615 | 100.000 | |
| 25-29 | Yes | 313 | 2366080 | 189219 | 3.8435 | 0.2729 | 1.5400 | 25.5809 | 1.5872 |
| | No | 983 | 6883314 | 377552 | 11.1813 | 0.5279 | 2.1441 | 74.4191 | 1.5872 |
| | Total | 1296 | 9249394 | 467221 | 15.0248 | 0.6057 | 2.1956 | 100.000 | |
| 30-34 | Yes | 275 | 2234545 | 188101 | 3.6298 | 0.2797 | 1.7094 | 21.7527 | 1.4772 |
| | No | 1080 | 8037936 | 396369 | 13.0569 | 0.4906 | 1.6203 | 78.2473 | 1.4772 |
| | Total | 1355 | 10272481 | 477661 | 16.6867 | 0.5571 | 1.7059 | 100.000 | |
| 35-39 | Yes | 170 | 1431768 | 140897 | 2.3258 | 0.2393 | 1.9257 | 13.1922 | 1.2698 |
| | No | 1100 | 9421336 | 427176 | 15.3041 | 0.6189 | 2.2583 | 86.8078 | 1.2698 |
| | Total | 1270 | 10853104 | 441417 | 17.6299 | 0.6615 | 2.3026 | 100.000 | |
| 40-44 | Yes | 98 | 868678 | 98464 | 1.4111 | 0.1540 | 1.3032 | 7.5458 | 0.8347 |
| | No | 1111 | 10643329 | 625810 | 17.2892 | 0.7818 | 3.2666 | 92.4542 | 0.8347 |
| | Total | 1209 | 11512007 | 647860 | 18.7002 | 0.7914 | 3.1481 | 100.000 | |
| Total | Yes | 1467 | 11662345 | 590372 | 18.9445 | 0.6579 | 2.1540 | | |
| | No | 6176 | 49898370 | 1489826 | 81.0555 | 0.6579 | 2.1540 | | |
| | Total | 7643 | 61560715 | 1873490 | 100.000 | | | | |

**Figure 11. SAS 9.1 output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002**

This program defines a data set called NSFG.EX1, which is derived from the NSFG female file. Two new variables are added to the female file in NSFG.EX1: an age group variable that defines 5-year age groups, and the outcome variable ''pill use'' which is equal to one if the value of CONSTAT1, the current contraceptive status, is equal to 6. PROC SURVEYFREQ is then applied to NSFG.EX1, and incorporates three survey design features in estimation: the stratum variable is identified as SEST, the cluster variable as SECU, and the weight as FINALWGT. The table specification cross-classifies the age group variable agerx by pill use, and requests that row percentages be estimated and design effect calculated for those row percentages.

This program generated the output presented in figure 11. The first panel is standard PROC SURVEYFREQ output that does not contain the estimated percentages of interest. The estimated row percentages and their standard errors are given in the second panel. The estimated percent using the oral contraceptive pill is highest for women 20–24 years of age, 31.7826 percent, with estimated standard error of 1.9966.

The SUDAAN version of the same program to compute similar estimates can be computed using a cross-tabulation program as well, PROC CROSSTAB (see figure 12). The DATA step illustrated for PROC SURVEYFREQ can be used for SUDAAN as well. In SUDAAN, the data must be sorted by stratum and by sampling error computing unit within stratum.

The PROC CROSSTAB command includes several specifications:

1. The input data set, NSFG.EX1
2. A request for estimation of design effects, DEFF
3. A design specification, DESIGN = WR.

The appropriate specification for the NSFG Cycle 6 data is DESIGN=WR, with a replacement variance estimation

**SAS 9.1 Output cont.**

```
% Females using oral contraceptive pill                          2

        The SURVEYFREQ Procedure

         Table of AGERX by PILL

                          Row     Std Err of
AGERX      PILL        Percent    Row Percent
──────────────────────────────────────────────
15-19       Yes        16.6155         1.4964
             No        83.3845         1.4964

          Total        100.000
----------------------------------------------
20-24       Yes        31.7826         1.9966
             No        68.2174         1.9966

          Total        100.000
----------------------------------------------
25-29       Yes        25.5809         1.5872
             No        74.4191         1.5872

          Total        100.000
----------------------------------------------
30-34       Yes        21.7527         1.4772
             No        78.2473         1.4772

          Total        100.000
----------------------------------------------
35-39       Yes        13.1922         1.2698
             No        86.8078         1.2698

          Total        100.000
----------------------------------------------
40-44       Yes         7.5458         0.8347
             No        92.4542         0.8347

          Total        100.000
----------------------------------------------
Total       Yes
             No

          Total
──────────────────────────────────────────────
```

**Figure 11—Continued**

procedure within each stratum, appropriate for the NSFG under the assumptions of ultimate cluster sampling.

The NEST statement is used in SUDAAN to specify the stratification and clustering in one statement. The first variable in the next statement, SEST, is treated as the stratum variable. The second is treated as the cluster within the stratum, SECU. As in PROC SURVEYFREQ, the WEIGHT statement specifies the weight variable FINALWGT.

The PROC CROSSTAB specification of the table is somewhat more complicated than PROC SURVEYFREQ. A specification is given for each variable in the cross tabulation of the number of categories in each variable. The SUBGROUP statement identifies the cross-tabulation variables, while the LEVEL statement gives the number of categories in each (6 for AGERX and 2 for PILL). The TABLE statement combines the two variables to produce a cross-tabulation of age group as rows, and contraceptive use as columns.

SUDAAN has a wide variety of printed estimates that can be produced. The PRINT specification includes the unweighted count of cases in each cell (NSUM), the sum of weights for cases in each cell (WSUM), the row percentages (ROWPER), the standard error of the row percentages (SEROW), and the design effect of the row percentages (DEFFROW). There are numerous other print specifications that could be applied, but these in SUDAAN will print the same estimates obtained above in PROC SURVEYFREQ.

The output from this SUDAAN run is presented in figure 13. SUDAAN includes the weighted and unweighted counts in each cell and estimated row percentages and standard errors that are identical to those produced in SAS (the standard errors differ only in the third significant digit).

STATA uses a much more compact code specification than SAS or SUDAAN and the output is also considerably more compact (see figure 14). STATA has a set of "survey" commands that are preceded by the abbreviation "svy," which allow many standard analyses to handle the complex sample design features. In STATA, one must specify the data set in a USE statement before generating new variables, and specify the survey design features through a SVYSET statement before presenting the requested SVY command.

**SUDAAN Program**

```
(same recode as required in SAS 9.1)

proc sort data=NSFG.EX1;
by SEST SECU_R;
proc crosstab data=NSFG.EX1 design=wr deff;
nest SEST SECU_R;
weight FINALWGT;
subgroup agerx pill;
levels 6 2;
table agerx * pill;
print nsum wsum rowper serow deffrow;
run;
```

**Figure 12. SUDAAN program code for illustration 1**

**SUDAAN 8.0.2 Output**

```
                    S U D A A N
   Software for the Statistical Analysis of Correlated Data
 Copyright      Research Triangle Institute      January 2003
                      Release 8.0.2

Number of observations read    :   7643    Weighted count : 61560715
Denominator degrees of freedom :     84

Variance Estimation Method: Taylor Series (WR)
by: AGERX, using the pill
```

| AGERX | | Using the pill: Total | Yes | No |
|---|---|---|---|---|
| Total | Sample Size | 7643.0000 | 1467.0000 | 6176.0000 |
| | Weighted Size | 61560714.7761 | 11662344.8777 | 49898369.8984 |
| | Row Percent | 100.0000 | 18.9445 | 81.0555 |
| | SE Row Percent | 0.0000 | 0.6579 | 0.6579 |
| | DEFF Row Percent #4 | . | 2.1543 | 2.1543 |
| 15-19 | Sample Size | 1150.0000 | 187.0000 | 963.0000 |
| | Weighted Size | 9834108.6926 | 1633985.7873 | 8200122.9053 |
| | Row Percent | 100.0000 | 16.6155 | 83.3845 |
| | SE Row Percent | 0.0000 | 1.4964 | 1.4964 |
| | DEFF Row Percent #4 | . | 1.8587 | 1.8587 |
| 20-24 | Sample Size | 1363.0000 | 424.0000 | 939.0000 |
| | Weighted Size | 9839619.5662 | 3127289.0363 | 6712330.5299 |
| | Row Percent | 100.0000 | 31.7826 | 68.2174 |
| | SE Row Percent | 0.0000 | 1.9966 | 1.9966 |
| | DEFF Row Percent #4 | . | 2.5061 | 2.5061 |
| 25-29 | Sample Size | 1296.0000 | 313.0000 | 983.0000 |
| | Weighted Size | 9249394.2563 | 2366079.9438 | 6883314.3125 |
| | Row Percent | 100.0000 | 25.5809 | 74.4191 |
| | SE Row Percent | 0.0000 | 1.5872 | 1.5872 |
| | DEFF Row Percent #4 | . | 1.7150 | 1.7150 |
| 30-34 | Sample Size | 1355.0000 | 275.0000 | 1080.0000 |
| | Weighted Size | 10272481.3018 | 2234545.0246 | 8037936.2773 |
| | Row Percent | 100.0000 | 21.7527 | 78.2473 |
| | SE Row Percent | 0.0000 | 1.4772 | 1.4772 |
| | DEFF Row Percent #4 | . | 1.7371 | 1.7371 |
| 35-39 | Sample Size | 1270.0000 | 170.0000 | 1100.0000 |
| | Weighted Size | 10853103.9617 | 1431767.5693 | 9421336.3924 |
| | Row Percent | 100.0000 | 13.1922 | 86.8078 |
| | SE Row Percent | 0.0000 | 1.2698 | 1.2698 |
| | DEFF Row Percent #4 | . | 1.7881 | 1.7881 |
| 40-44 | Sample Size | 1209.0000 | 98.0000 | 1111.0000 |
| | Weighted Size | 11512006.9975 | 868677.5165 | 10643329.4810 |
| | Row Percent | 100.0000 | 7.5458 | 92.4542 |
| | SE Row Percent | 0.0000 | 0.8347 | 0.8347 |
| | DEFF Row Percent #4 | . | 1.2075 | 1.2075 |

**Figure 13. SUDAAN 8.0.2 output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002**

The USE statement specifies that the female data set (females.dta, in STATA) be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU) variables to be used in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command, or running *svyset* again with different settings. The *generate* and *replace* statements create the recodes of "agerx" and "pill." The syntax is somewhat different than SAS. For example, for "agerx," first a variable "agerx" is *generated* with an assigned value. Then the value is replaced if the AGER is in an age range other than 15–19. Finally, the *svytab* command produces a cross-tabulation of "agerx" and "pill," and adds specifications for row percentages (*row*), design effects for the row percentages (*deff percent*), and standard errors of the estimated row percentages (*se*). *Svytab* uses the design specifications given by *svyset*.

The STATA output is shown in figure 15. In addition to the percentages and standard errors in the cross tabulation, STATA produces several test statistics for the hypothesis of no difference in proportions across the rows. The F-test takes into account the complex sample design, while the uncorrected chi-square test does not. Also given is a generalized design effect that can be used in adjusting chi-square tests of independence for cross-tabulated data.

The last program illustrated is WesVar, a stand-alone menu-driven program. WesVar uses a series of menus to specify data preparation, recodes, and estimates. The presentation of the screens showing the various menus and steps required to generate a WesVar format data set, prepare variables for analysis, and specify cross tabulations are quite extensive. Only one screen image is given in this report (see figure 16). The full set of screens is given at the NSFG Cycle 6 Web site, www.cdc.gov/nchs/about/major/nsfg/nsfgvar.htm, together with the output from WesVar's cross-tabulation of age group by pill use.

The remaining illustrations in Appendixes IV–VII, and on the Web site, provide additional examples of the use of these four packages for NSFG Cycle 6 data analysis. They show recode, table specification, analysis options, and other features not shown in this first illustration. NSFG Cycle 6 users not familiar with the use of this software are urged to study these examples to prepare programs for other analyses of interest.

Users of the NSFG Cycle 6 may follow a strategy that uses these software systems only at a last stage of estimation. Many users will compute estimates and statistical tests using the FINALWGT using standard statistical software procedures in SPSS, SAS, STATA, and other systems. When they have produced the final, or near final table or analysis, they then use one of these packages to compute standard errors and estimated values, taking into

**STATA 8.0 Program**

```
use "EX1.DTA"

svyset [pweight=FINALWGT], strata(SEST) psu(SECU_R)

generate agerx=1 if AGER <=19
replace  agerx=2 if AGER >=20 & AGER <=24
replace  agerx=3 if AGER >=25 & AGER <=29
replace  agerx=4 if AGER >=30 & AGER <=34
replace  agerx=5 if AGER >=35 & AGER <=39
replace  agerx=6 if AGER >=40

generate pill=2
replace pill=1 if CONSTAT1==6

svytab agerx pill, row se deff percent
```

**Figure 14. STATA 8.0 program code for illustration 1**

**STATA 8.0 Output**

```
pweight:  finalwgt                          Number of obs      =      7643
Strata:   sest                              Number of strata   =        84
PSU:      secu_r                            Number of PSUs     =       168
                                            Population size    =  61560715


          |       Using the pill
          |
    agerx |     Yes        No     Total
----------+-------------------------------
    15-19 |   16.62     83.38       100
          |  (1.496)   (1.496)
          |   66.23     14.82
          |
    20-24 |   31.78     68.22       100
          |  (1.997)   (1.997)
          |   63.18     31.36
          |
    25-29 |   25.58     74.42       100
          |  (1.587)   (1.587)
          |   52.09     19.39
          |
    30-34 |   21.75     78.25       100
          |  (1.477)   (1.477)
          |   47.67     14.69
          |
    35-39 |   13.19     86.81       100
          |   (1.27)    (1.27)
          |   54.24     9.506
          |
    40-44 |   7.546     92.45       100
          |  (.8347)   (.8347)
          |   38.28     3.724
          |
    Total |   18.94     81.06       100
          |  (.6579)   (.6579)
          |   2.154     2.154

  Key:  row percentages
        (standard errors of row percentages)
        deff for variances of row percentages

  Pearson:
    Uncorrected   chi2(5)          =  324.8924
    Design-based  F(4.63, 388.69)  =   36.6663     P = 0.0000

  Mean generalized deff            =    2.0255
  CV of generalized deffs          =    0.5664
```

**Figure 15. STATA output from illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002**

account the full complexity of the sample design. Thus, they are able to operate for nearly all of their analytical work in a more familiar software environment before turning to the use of one of these packages designed to handle the complex sample design features such as weights, stratification, and clustered sample selection.

Finally, Appendix VIII provides estimates of design effects for variables from the female, male, and pregnancy files. Some NSFG Cycle 6 users may find these summary design effects useful in assessing the importance of using the sampling estimate software illustrated here, in the appendixes, and on the NSFG Web site.

# References


1. French DK. National Survey of Family Growth, Cycle 1: Sample design, estimation procedures, and variance estimation. Vital Health Stat 2(76). 1978.
2. Grady WR. National Survey of Family Growth, Cycle 2: Sample design, estimation procedures, and variance estimation. Vital Health Stat 2(87). 1981.
3. Bachrach CA, Horn MC, Mosher WD, Shimizu I. National Survey of Family Growth, Cycle 3: Sample design, weighting, and variance estimation. Vital Health Stat (2)98. 1985.
4. Judkins DR, Mosher WD, Botman S. National Survey of Family Growth: Design, estimation, and inference. Vital Health Stat (2)109. 1991.
5. Potter FJ, Iannacchione VG, Mosher WD, Mason RE, Kavee JD. Sample design, sampling weights, imputation, and variance estimation in the 1995 National Survey of Family Growth. Vital Health Stat (2)124. 1998.
6. Kelly JE, Mosher WD, Duffer AD, Jr, and Kinsey SH. Plan and operation of the 1995 National Survey of Family Growth. Vital Health Stat (1)36. 1997.
7. Kish L, Scott A. Retaining units after changing strata and probabilities. J Am Stat Assoc 66:461–70. 1971.
8. Hansen MH, Hurwitz WN. 1946. The problem of nonresponse in sample surveys. J Am Stat Assoc 41:517–29. 1946.
9. Groves RM, Benson G, Mosher WD, Rosenbaum J, et al. Plan and operation of Cycle 6 of the National Survey of Family Growth. Vital Health Stat (1)42. 2005.

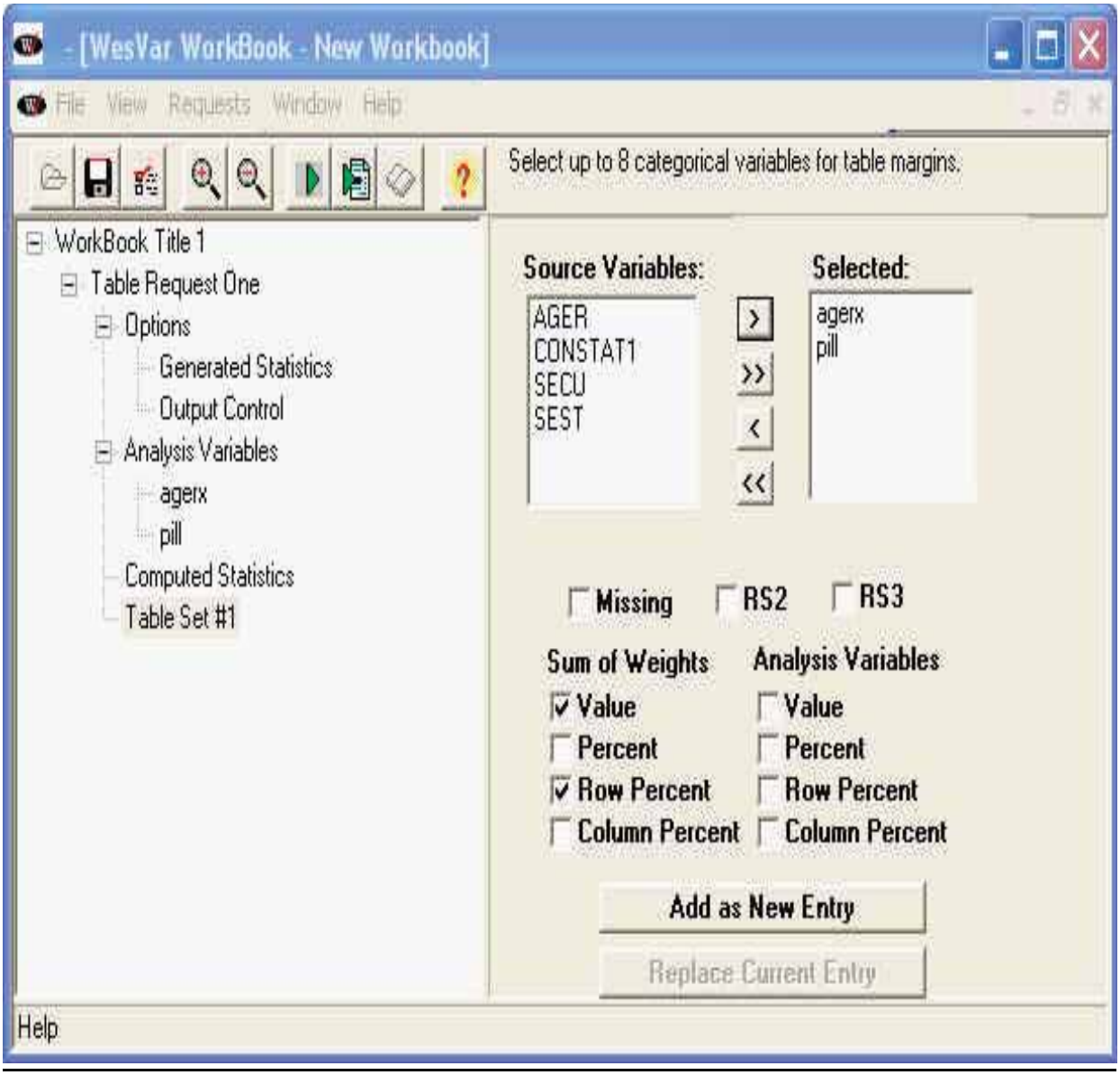


**Figure 16. Wesvar table specification window for illustration 1 for estimation of the percent of women using the oral contraceptive pill in each of six 5-year age groups, National Survey of Family Growth Cycle 6, 2002**


10. Groves RM, Couper MP. Nonresponse in Household Surveys. New York: John Wiley and Sons. 1998.
11. American Association for Public Opinion Research. Standard Definitions: Final dispositions of case codes and outcome rates for surveys. Available at: http://www.aapor.org. On-line edition 3.1. Revised 2005.
12. Groves RM, Dillman DA, Eltinge JL, Little RJA, editors. Survey Nonresponse, New York: Wiley and Sons. 2001.
13. Kalton G, Kasprzyk D. The treatment of missing survey data. Survey Methodology 12: 1–16. 1986.
14. Raghunathan TE, Lepkowski JM, Van Hoewyk J, Solenberger P. Sequential Regression Imputation for Survey Data. Survey Methodology 27:85–96. 2001.
15. Wolter KM. Introduction to variance estimation. New York: Springer-Verlag. 1985.
16. Kish L, Frankel MR. Inference from complex samples. J Royal Statistical Soc B 36:1–37. 1974.
17. Mahalanobis PC. "On large-scale sample surveys," Philosophical transactions of the Royal Soc B 231:329–451. 1944.
18. Deming WE. On simplifications of sampling design through replication with equal probabilities and without stages. J Am Stat Assoc 51:24–53. 1956.
19. McCarthy PJ. Replication. An approach to the analysis of data from complex sample surveys. Vital Health Stat (2)14. Washington, DC: U.S. Government Printing Office. 1996.
20. Rust K. Variance estimation for complex estimators in sample surveys. J Official Stat 1:381–97. 1985.

# Appendix I

## Definition of Terms

*ACASI*—Audio computer-assisted self-interviewing, in which the survey respondent uses a laptop computer to complete the last part of the NSFG questionnaire (Section K of the Male Questionnaire and Section J of the female questionnaire). The interviewer asks the respondent to use earphones to hear audio-recorded questions played over the laptop computer sound system through the earphones. The question text is also displayed on the laptop monitor. The respondent enters his or her answers using the computer keyboard. The respondent performed these steps out of the sight of the interviewer to increase the privacy of the answers.

*Blaise®*—A software system for computer-assisted questionnaires. Blaise® was developed by Statistics Netherlands and is licensed in the United States and Canada by Westat, Inc. When appropriately programmed, Blaise® can present the questions to the interviewer, route the respondent to the next appropriate question, record the answers entered by the interviewer, and check those answers for consistency with previous answers. Blaise® was used for Cycle 6 of the NSFG and an earlier version was used in Cycle 5. Further information is available at http://www.westat.com/statistical_software/blaise/.

*CAPI*—Computer-assisted personal interviewing, in which an interviewer uses a laptop computer to conduct the interview. The laptop displays the questions for the interviewer to ask the respondent and provides any necessary instructions the interviewer may need. Interviewers record the respondent's answers using the keyboard. Software directs the interviewer to the next appropriate question based on the answers entered.

*Contact rate*—The percentage of sample households where an interviewer talked with someone at the household at the screener stage (the screener contact rate); at the main interview stage, the percentage of sample persons who met with the interviewer on one or more visits to the household by the interviewer (the main interview contact rate).

*Cooperation rate*—The percentage of sample households who were contacted and granted a screener interview (screener cooperation rate); the percentage of sample persons contacted who granted a main interview (main interview cooperation rate).

*Coverage error*—Deviations between the characteristics (for example, values of estimated population characteristics) of the sampling frame and the desired target population. Coverage errors arise from the failure to list on the sampling frame (a) some households containing eligible persons and (b) some eligible persons within sample households.

*Domain*—A stratum; a group of sampling units (such as blocks) placed in the same subset from which a sample of units was selected.

*Double sample (or two-phase sample)*—A sample of a survey sample (hence the term, "double sample"). In NSFG Cycle 6, a sample of nonrespondent cases was selected after the completion of the first two phases of data collection, at the end of January 2003. The sample of nonrespondents was then released in Phase 3 for more intensive followup and interviewing efforts in February and early March 2003. The double sample was used in the NSFG Cycle 6 in an attempt to raise response rates and reduce nonresponse bias.

*Eligible household*—A household containing at least one person who was eligible for the NSFG Cycle 6—that is, males or females 15–44 years of age at the date on which the screener was completed in 2002, and living in the household population of the United States (all 50 States and the District of Columbia). Whether a selected household had an eligible person was not known until the household screener was conducted. If a household had two or more persons 15–44 years of age, one person was selected randomly.

*Eligibility rate*—The percentage of sample cases that are members of the target population. In NSFG, Cycle 6 the eligibility rate is the percentage of households that contain a person 15–44 years of age.

*Epsem*—Equal probability selection method; a sample design that gives all sample units an equal chance of selection.

*Institute for Social Research, University of Michigan*—The Institute for Social Research (ISR) at the University of Michigan conducted the fieldwork and data processing for Cycle 6 of the NSFG under a contract with NCHS. ISR has several centers that participated in the NSFG: the Survey Research Center provided overall coordination and was responsible for data collection, weighting, and variance estimation; the InterUniversity Consortium on Political and Social Research processed data and developed documentation and web-based systems; and the Population Studies Center provided substantive expertise on demography and family growth.

*Item imputation*—The process of assigning answers to cases with missing data ("don't know," "refused," or "not ascertained"). In the NSFG, item imputation was only performed on approximately 350 "recoded variables," or "recodes" (defined below, under "recodes"). The purposes of imputation are to make the data more complete, more consistent, easier to use, and, most importantly, to reduce bias caused by differential failure to respond. For example, if a respondent's educational level was missing and a value of "high school graduate" was assigned, her education was imputed. Imputation was done in two ways in Cycle 6 of the NSFG: logical and regression imputation. Logical imputation uses a subject-matter expert to assign a value based on the value of other variables for the case with missing data. Regression imputation uses a regression equation to estimate a value for a case with missing data. For most of the recoded variables for which imputation was done in the NSFG, less than 2 percent of the cases received an imputed value.

*Main interview*—An interview sought within sample households known to contain a member of the target population. If the screening interview reveals that the household is eligible for the survey, a main interview is requested. In NSFG Cycle 6, if the household had one or more members

15–44 years of age, one such person was selected at random for the main interview request.

*Measure of size*—A value assigned to every sampling unit in a sample selection. Typically measures of size are a count of units associated with the elements to be selected. For example, measures of size for NSFG PSUs could have been the count of occupied housing units obtained in the 2000 Census of Population and Housing, since sample selection within the PSU would have selected housing units. In NSFG Cycle 6, the counts of occupied housing units from the year 2000 Census were not yet available at the time of sample selection. Instead, an approximate measure of size was calculated for each PSU based on population counts from the year 2000 Census divided by an estimated average household size.

Measures of size were also used in the selection of eligible persons within the household (see figure 5) to increase the chances of selection of such groups as teenagers 15–19 years of age and females. Each person in the household was assigned a measure of size between 0 and 1, where the measures were predetermined values for each of 18 groups defined by age (3 groups), gender (2 groups), and race or ethnicity (3 groups). The measures of size were cumulated across eligible persons, a random number from zero to the sum of the measures was generated, and an individual was selected based on the cumulated measures of size.

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

*Paradata*—Information collected via computer software or interviewer observations describing the sample unit, interactions with sample household members, or features of the interview situation. NSFG Cycle 6 used observations of characteristics of sample housing units to reduce number of callbacks, statements made by respondents to diagnose their concerns, call record data to model response propensities, and observations of the respondent during ACASI for measurement error modeling. Some paradata are labeled as "process data."

*Phase*—A period of data collection during which the same set of sampling frame, mode of data collection, sample design, recruitment protocols, and measurement conditions are used. In NSFG Cycle 6 there were three phases: first, a quarter-sample in which interviewers were permitted to visit households as many times as they thought was necessary (March–May 2002); second, the full sample with prespecified call rules (May 2002–January 2003), and third, a subsample of nonrespondents from Phases 1–2 offered higher incentives by a subset of high-productivity interviewers (February 2003).

*Pretest*—A survey done before a main survey. It is usually smaller than the main survey. Its purpose is to test questionnaires, procedures, equipment, and personnel before they are used in the larger main survey.

*Primary sampling unit (PSU)*—The first-stage selection unit in a multistage area probability sample. In Cycle 6 of the NSFG, PSUs are counties or groups of counties in the United States; there were 121 PSUs selected into the NSFG sample selected from the more than 3,000 counties in the United States. Secondary sampling units—called segments and tertiary units—or selected households are drawn from the primary sampling units.

*Race or ethnicity*—Race or ethnicity is used in this report *as it was used to select the NSFG Cycle 6 sample*. Three categories were used for purposes of sample design: Hispanic, non-Hispanic black, and other. Hispanic and non-Hispanic black men and women were selected at higher rates than others in Cycle 6 of the NSFG to obtain adequate numbers of Hispanic and black persons to make reliable national estimates for these groups. Thus, when this report contains tables showing "race or ethnicity," the three categories are **those used to design and select the sample.** In contrast, in reports that are designed to present substantive results, the "other" category is often split into "non-Hispanic white" and "non-Hispanic other" categories, and those who reported only one race are shown separately from the small percentage who reported more than one race.

*Receptor case*—In imputation, a case with missing data that receives the imputed value from another (donor) case. If Case B has missing data and Case B receives an imputed value from Case A, then Case A is the donor case and Case B is the receptor case.

*Recodes or recoded variables*—It was not possible to edit or impute all of the variables in the NSFG data file. NSFG staff selected about 350 variables from the NSFG data file that were to be constructed, edited, and imputed. These were called recodes or recoded variables. Recodes were variables that were likely to be used frequently by NCHS and other data users. They were edited for consistency, and missing values were imputed. Many (but not all) of these recoded variables are constructed from other variables in the NSFG; some are constructed from a large number of other variables. Other variables in the data file were not edited or imputed in this way.

*Replicate*—A probability subsample of the full sample design. The complete sample consists of several replicate subsamples, each of which is a small national sample of housing units. Replicate samples are released over the data collection to control the workflow of the interviewers. In responsive designs, such as the NSFG Cycle 6, early replicates are used to measure key cost and error features of a survey.

*Response rate*—The number of respondents to a survey divided by the number of eligible persons in the sample. In this report, the response rate is the number of respondents (15–44 years of age) divided by the number of

eligible persons (15–44 years of age). Given that not all screeners were completed, the number of eligible persons is not known precisely, so the number of eligible persons was estimated.

*Respondent*—A person who answers, or responds to, a survey. In Cycle 6 of the NSFG, the "respondents" are the 12,571 men and women 15–44 years of age who completed the NSFG interview.

*Responsive design*—Survey designs that

- Pre-identify a set of design features potentially affecting costs and errors of survey statistics
- Identify a set of indicators of the cost and error properties of those features
- Monitor those indicators in initial phases of data collection
- Alter the active features of the survey in subsequent phases based on cost/error tradeoff decision rules
- Combine data from the separate design phases into a single estimator

A responsive design approach was used in Cycle 6 of the NSFG to maximize response rates and reduce bias while controlling costs.

*Sample line*—A housing unit listed within a sample segment prior to the data-collection phase of the NSFG, which was subsequently sampled for inclusion in the NSFG sample.

*Sampling variance*—The sampling variance is a measure of the variation of a statistic, such as a proportion or a mean, that is due to having taken a random sample instead of collecting data from every person in the full population. It measures the variation of a statistic such as the estimated proportion or mean over repeated samples. The sampling variance is zero when the full population is observed, as in a census. For the NSFG, the sampling variance estimate is a function of the sampling design and the population parameter being estimated (for example, a proportion or a mean). Many common statistical software packages compute "population" variances, which may underestimate the sampling variance. Estimating the sampling variance requires special software, such as those discussed in Appendixes IV, V, VI, and VII.

*Sampling weight*—For a respondent in the NSFG, the number of persons in the target population that he or she represents. For example, if a man in the sample represents 12,000 men in his age and race or ethnicity category, then his "sampling weight" is 12,000. The NSFG Cycle 6 sampling weight adjusts for different sampling rates (of the age and race or ethnicity groups), different response rates, and different coverage rates among persons in the sample, so that accurate national estimates can be made from the sample. Because it adjusts for all these factors, it is sometimes called a "fully adjusted" sampling weight.

*Screening interview*— (Sometimes called a "screener") a (usually short) set of questions, asked of one adult in a household, with the chief goal of determining whether the household is eligible for the chosen target population of the survey. In the NSFG Cycle 6, the screening interview consisted of a household roster collecting age, race, and sex. Households having one or more persons 15–44 years of age were eligible for a main interview.

*Self-representing area*—A county or group of counties selected with certainty; a self-representing area forms a primary sampling unit with population counts large enough to be equal to or greater than the typical stratum size in the U.S. national sample. Such PSUs are thus represented in all draws of a national sample using the design. The sampling probabilities for persons in such areas are designed to be equal to that applicable in smaller PSUs, called non-self-representing areas. When PSUs in a sampling frame are "selected with certainty," it means that they are automatically included in the sample. For Cycle 6 of the NSFG, 28 of the 121 PSUs were selected with certainty and are referred to as "self-representing" PSUs.

*Simple random sample*—A sample in which all members of the population are selected directly and have an equal chance to be selected for the sample. The NSFG is not a simple random sample. The NSFG sample was stratified, selected in stages, and employed unequal chances of selection for the respondents, varying by age, race or ethnicity, and sex. Such designs are referred to as "complex" and require special software to account completely for all features of the design.

*Strata; Stratification*—Stratification is the partitioning of a population of sampling units into mutually exclusive categories (strata). Typically, stratification is used to increase the precision of survey estimates for subpopulations important to the survey's objectives. In Cycle 6 of the NSFG, those groups include teenagers (15–19 years of age), Hispanic men and women, and non-Hispanic black men and women. To obtain larger and more reliable samples of these groups, the NSFG sample was stratified: in the first stage of selection, PSUs were stratified using socioeconomic and demographic variables; in the second stage of selection, segments within each PSU were stratified by the concentration of black and Hispanic populations.

*SurveyTrak*—A software-based sample administration system used in Cycle 6 of the NSFG. The system is used by interviewers on laptop computers to document their sample assignment, to organize the activities of their workday, to prompt them for appointments to be kept, to record results of each call attempt, to record observations of the sample housing unit, and in all other ways to keep track of their job duties.

*Target population*—The population to be described by estimates from the survey. In NSFG Cycle 6 the target population was the household population of the United States. It omits persons in prisons, hospitals, and other large residences under central control. College students living in dormitories were interviewed but sampled through their parents' or guardians' households. In addition, military personnel living in households off military bases were included in the sample.

*WEBDOC*—An electronic "codebook" for Cycle 6 of the NSFG, presenting data items, frequencies, code categories, and other survey documentation. The NSFG WEBDOC can be found at http://www.cdc.gov/nchs/about/major/nsfg/nsfgdoc.htm.

*Weight*—see "Sampling weight."

# Appendix II

## Variables Imputed and Types of Imputation for Female, Male, and Pregnancy Files

**Table I. Imputed case counts in the female file**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| ABORTION | 4,945 | 0 | 31 | 0.6 |
| ADDEXP | 7,643 | 30 | 73 | 1.3 |
| ADVICE | 569 | 7 | 0 | 1.2 |
| AGEBABY1 | 4,413 | 20 | 31 | 1.1 |
| AGECON01 | 5,033 | 42 | 0 | 0.8 |
| AGECON02 | 3,766 | 38 | 0 | 1.0 |
| AGECON03 | 2,334 | 28 | 0 | 1.2 |
| AGECON04 | 1,224 | 27 | 0 | 2.2 |
| AGECON05 | 613 | 24 | 0 | 3.8 |
| AGECON06 | 308 | 19 | 0 | 5.8 |
| AGECON07 | 158 | 17 | 0 | 9.7 |
| AGECON08 | 78 | 12 | 0 | 13.3 |
| AGECON09 | 38 | 5 | 0 | 11.6 |
| AGECON10 | 17 | 2 | 0 | 10.5 |
| AGECON12 | 5 | 1 | 0 | 16.7 |
| AGECON13 | 3 | 1 | 0 | 25.0 |
| AGEDD1 | 1,290 | 282 | 1 | 18.0 |
| AGEDISS1 | 1,550 | 502 | 0 | 24.5 |
| AGEMOMB1 | 7,643 | 51 | 0 | 0.7 |
| AGEPRG01 | 4,945 | 40 | 1 | 0.8 |
| AGEPRG02 | 3,662 | 36 | 2 | 1.0 |
| AGEPRG03 | 2,253 | 24 | 4 | 1.2 |
| AGEPRG04 | 1,186 | 21 | 5 | 2.1 |
| AGEPRG05 | 596 | 17 | 7 | 3.9 |
| AGEPRG06 | 300 | 13 | 6 | 6.0 |
| AGEPRG07 | 147 | 11 | 5 | 9.8 |
| AGEPRG08 | 76 | 7 | 5 | 13.6 |
| AGEPRG09 | 36 | 4 | 1 | 12.2 |
| AGEPRG10 | 17 | 0 | 2 | 10.5 |
| AGEPRG12 | 5 | 1 | 0 | 16.7 |
| AGEPRG13 | 3 | 1 | 0 | 25.0 |
| ANYMSCHP | 7,172 | 7 | 0 | 0.1 |
| ANYPRGHP | 7,172 | 7 | 0 | 0.1 |
| B1PREMAR | 4,413 | 31 | 31 | 1.4 |
| BIRTHS5 | 4,945 | 0 | 31 | 0.6 |
| CEBOW | 4,741 | 0 | 31 | 0.6 |
| CEBOWC | 4,741 | 0 | 31 | 0.6 |
| COH1DUR | 3,904 | 148 | 12 | 3.9 |
| COHAB1 | 3,904 | 103 | 12 | 2.9 |
| COHOUT | 3,904 | 103 | 57 | 3.9 |
| COHSTAT | 7,643 | 49 | 12 | 0.8 |
| COMPREG | 7,643 | 0 | 7 | 0.1 |
| CON1MAR1 | 5,531 | 66 | 0 | 1.2 |
| CONDOMR | 6,785 | 5 | 0 | 0.1 |
| CONSTAT1 | 7,643 | 55 | 0 | 0.7 |
| CONSTAT2 | 7,643 | 55 | 0 | 0.7 |
| CONSTAT3 | 7,643 | 55 | 0 | 0.7 |
| CONSTAT4 | 7,643 | 55 | 0 | 0.7 |
| DATBABY1 | 4,413 | 8 | 31 | 0.9 |
| DATCON01 | 5,033 | 28 | 1 | 0.6 |
| DATCON02 | 3,766 | 27 | 2 | 0.8 |
| DATCON03 | 2,334 | 18 | 4 | 0.9 |
| DATCON04 | 1,224 | 19 | 4 | 1.8 |
| DATCON05 | 613 | 18 | 5 | 3.6 |
| DATCON06 | 308 | 14 | 4 | 5.5 |
| DATCON07 | 158 | 13 | 4 | 9.7 |
| DATCON08 | 78 | 11 | 1 | 13.3 |
| DATCON09 | 38 | 4 | 1 | 11.6 |
| DATCON10 | 17 | 2 | 0 | 10.5 |
| DATCON12 | 5 | 1 | 0 | 16.7 |

**Table I. Imputed case counts in the female file—Con.**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| DATCON13 | 3 | 1 | 0 | 25.0 |
| DATEND01 | 4,945 | 27 | 1 | 0.6 |
| DATEND02 | 3,662 | 27 | 2 | 0.8 |
| DATEND03 | 2,253 | 18 | 4 | 1.0 |
| DATEND04 | 1,186 | 19 | 3 | 1.8 |
| DATEND05 | 596 | 19 | 4 | 3.7 |
| DATEND06 | 300 | 14 | 4 | 5.7 |
| DATEND07 | 147 | 13 | 3 | 9.8 |
| DATEND08 | 76 | 11 | 1 | 13.6 |
| DATEND09 | 36 | 4 | 1 | 12.2 |
| DATEND10 | 17 | 2 | 0 | 10.5 |
| DATEND12 | 5 | 1 | 0 | 16.7 |
| DATEND13 | 3 | 1 | 0 | 25.0 |
| DATESEX1 | 6,776 | 36 | 129 | 2.4 |
| DATEUSE1 | 6,631 | 101 | 0 | 1.5 |
| DD1REMAR | 1,290 | 283 | 3 | 18.1 |
| EDUCAT | 7,643 | 3 | 0 | 0.0 |
| EDUCMOM | 7,643 | 99 | 0 | 1.3 |
| ENDOMET | 569 | 7 | 0 | 1.2 |
| EVHIVTST | 7,643 | 53 | 0 | 0.7 |
| FIBROIDS | 569 | 7 | 0 | 1.2 |
| FMAR1AGE | 4,126 | 48 | 0 | 1.1 |
| FMARNO | 7,643 | 8 | 0 | 0.1 |
| FMETHOD1 | 6,764 | 18 | 0 | 0.3 |
| FPREGMED | 599 | 1 | 0 | 0.2 |
| FPTIT12 | 1,137 | 42 | 0 | 3.6 |
| FPTITABO | 96 | 3 | 0 | 3.0 |
| FPTITBC | 2,682 | 35 | 0 | 1.3 |
| FPTITCBC | 1,526 | 27 | 0 | 1.7 |
| FPTITCEC | 288 | 9 | 0 | 3.0 |
| FPTITCHK | 1,879 | 22 | 0 | 1.2 |
| FPTITCST | 372 | 6 | 0 | 1.6 |
| FPTITEC | 93 | 2 | 0 | 2.1 |
| FPTITMED | 1,540 | 54 | 0 | 3.4 |
| FPTITPAP | 4,968 | 43 | 0 | 0.9 |
| FPTITPEL | 4,517 | 39 | 0 | 0.9 |
| FPTITPPR | 495 | 7 | 0 | 1.4 |
| FPTITPRE | 1,648 | 26 | 0 | 1.6 |
| FPTITPRN | 601 | 2 | 0 | 0.3 |
| FPTITSTD | 1,092 | 14 | 0 | 1.3 |
| FPTITSTE | 157 | 14 | 0 | 8.2 |
| FSEXPAGE | 6,785 | 60 | 0 | 0.9 |
| HIEDUC | 7,643 | 0 | 5 | 0.1 |
| HISPANIC | 7,643 | 3 | 0 | 0.0 |
| INFERT | 3,812 | 0 | 28 | 0.7 |
| INFERTH | 569 | 7 | 0 | 1.2 |
| INFERTR | 569 | 7 | 0 | 1.2 |
| INFEVER | 7,172 | 9 | 0 | 0.1 |
| INSEM | 569 | 7 | 0 | 1.2 |
| INSURANC | 7,643 | 54 | 6 | 0.8 |
| INTCTFAM | 7,643 | 0 | 0 | 0.0 |
| INTENT | 7,643 | 2 | 67 | 0.9 |
| INVITRO | 569 | 7 | 0 | 1.2 |
| LABORFOR | 7,643 | 5 | 0 | 0.1 |
| LBPREGS | 4,945 | 0 | 31 | 0.6 |
| LIFPRTNR | 7,643 | 75 | 1 | 1.0 |
| LIV1CHLD | 4,413 | 0 | 31 | 0.7 |
| LOSSNUM | 4,945 | 0 | 31 | 0.6 |
| LSEXDATE | 6,785 | 45 | 31 | 1.1 |
| LSEXRAGE | 6,785 | 92 | 0 | 1.3 |
| MAR1BIR1 | 5,213 | 38 | 31 | 1.3 |
| MAR1CON1 | 5,535 | 57 | 0 | 1.0 |
| MAR1DISS | 4,126 | 523 | 0 | 11.2 |
| MARCON01 | 5,033 | 176 | 0 | 3.4 |
| MARCON02 | 3,766 | 249 | 0 | 6.2 |
| MARCON03 | 2,334 | 207 | 0 | 8.1 |
| MARCON04 | 1,224 | 145 | 0 | 10.6 |
| MARCON05 | 613 | 100 | 0 | 14.0 |

**Table I. Imputed case counts in the female file—Con.**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| MARCON06 | 308 | 64 | 0 | 17.2 |
| MARCON07 | 158 | 40 | 0 | 20.2 |
| MARCON08 | 78 | 17 | 0 | 17.9 |
| MARCON09 | 38 | 7 | 0 | 15.6 |
| MARCON10 | 17 | 3 | 0 | 15.0 |
| MARCON11 | 8 | 1 | 0 | 11.1 |
| MARDAT01 | 4,126 | 37 | 0 | 0.9 |
| MARDAT02 | 690 | 4 | 0 | 0.6 |
| MARDIS01 | 1,550 | 509 | 0 | 24.7 |
| MARDIS02 | 258 | 135 | 0 | 34.4 |
| MARDIS03 | 46 | 29 | 0 | 38.7 |
| MARDIS04 | 9 | 6 | 0 | 40.0 |
| MARDIS05 | 2 | 2 | 0 | 50.0 |
| MAREND01 | 1,550 | 122 | 0 | 7.3 |
| MAREND02 | 258 | 37 | 0 | 12.5 |
| MAREND03 | 46 | 7 | 0 | 13.2 |
| MAREND04 | 9 | 1 | 0 | 10.0 |
| MAROUT01 | 4,945 | 224 | 0 | 4.3 |
| MAROUT02 | 3,662 | 273 | 0 | 6.9 |
| MAROUT03 | 2,252 | 213 | 0 | 8.6 |
| MAROUT04 | 1,186 | 150 | 0 | 11.2 |
| MAROUT05 | 596 | 102 | 0 | 14.6 |
| MAROUT06 | 300 | 65 | 0 | 17.8 |
| MAROUT07 | 147 | 38 | 0 | 20.5 |
| MAROUT08 | 76 | 17 | 0 | 18.3 |
| MAROUT09 | 36 | 6 | 0 | 14.3 |
| MAROUT10 | 17 | 3 | 0 | 15.0 |
| MAROUT11 | 8 | 1 | 0 | 11.1 |
| METH12M1 | 4,408 | 37 | 0 | 0.8 |
| METH12M2 | 626 | 37 | 0 | 5.6 |
| METH12M3 | 34 | 37 | 0 | 52.1 |
| METH12M4 | 6 | 37 | 0 | 86.0 |
| METH3M1 | 3,986 | 75 | 0 | 1.8 |
| METH3M2 | 531 | 75 | 0 | 12.4 |
| METH3M3 | 33 | 75 | 0 | 69.4 |
| METH3M4 | 6 | 75 | 0 | 92.6 |
| MTHUSE12 | 6,019 | 32 | 0 | 0.5 |
| MTHUSE3 | 5,426 | 79 | 0 | 1.4 |
| NOSEX12 | 7,643 | 80 | 61 | 1.8 |
| NUMP3MOS | 6,785 | 231 | 0 | 3.3 |
| OLDWP01 | 5,033 | 16 | 0 | 0.3 |
| OLDWP02 | 3,766 | 12 | 0 | 0.3 |
| OLDWP03 | 2,334 | 14 | 0 | 0.6 |
| OLDWP04 | 1,224 | 10 | 0 | 0.8 |
| OLDWP05 | 613 | 11 | 0 | 1.8 |
| OLDWP06 | 308 | 6 | 0 | 1.9 |
| OLDWP07 | 158 | 6 | 0 | 3.7 |
| OLDWP08 | 78 | 4 | 0 | 4.9 |
| OLDWR01 | 5,033 | 58 | 0 | 1.1 |
| OLDWR02 | 3,766 | 44 | 0 | 1.2 |
| OLDWR03 | 2,334 | 37 | 0 | 1.6 |
| OLDWR04 | 1,224 | 20 | 0 | 1.6 |
| OLDWR05 | 613 | 18 | 0 | 2.9 |
| OLDWR06 | 308 | 5 | 0 | 1.6 |
| OLDWR07 | 158 | 6 | 0 | 3.7 |
| OLDWR08 | 78 | 3 | 0 | 3.7 |
| OLDWR09 | 38 | 1 | 0 | 2.6 |
| OUTCOM01 | 5,033 | 0 | 1 | 0.0 |
| OUTCOM02 | 3,766 | 0 | 2 | 0.1 |
| OUTCOM03 | 2,334 | 0 | 4 | 0.2 |
| OUTCOM04 | 1,224 | 0 | 4 | 0.3 |
| OUTCOM05 | 613 | 0 | 8 | 1.3 |
| OUTCOM06 | 308 | 0 | 6 | 1.9 |
| OUTCOM07 | 158 | 0 | 5 | 3.1 |
| OUTCOM08 | 78 | 0 | 5 | 6.0 |
| OUTCOM09 | 38 | 0 | 1 | 2.6 |
| OUTCOM10 | 17 | 0 | 2 | 10.5 |
| OVULATE | 569 | 7 | 0 | 1.2 |

**Table I. Imputed case counts in the female file—Con.**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| PARAGE14 | 7,643 | 7 | 0 | 0.1 |
| PARITY | 7,643 | 0 | 31 | 0.4 |
| PARTS1YR | 6,785 | 19 | 0 | 0.3 |
| PIDTREAT | 7,643 | 9 | 1 | 0.1 |
| PILLR | 7,643 | 7 | 0 | 0.1 |
| POVERTY | 7,643 | 630 | 0 | 7.6 |
| PREGNUM | 7,643 | 0 | 7 | 0.1 |
| PUBASSIS | 7,643 | 156 | 0 | 2.0 |
| RELIGION | 7,643 | 23 | 0 | 0.3 |
| RMAROUT01 | 4,945 | 219 | 0 | 4.2 |
| RMAROUT02 | 3,662 | 263 | 0 | 6.7 |
| RMAROUT03 | 2,253 | 196 | 0 | 8.0 |
| RMAROUT04 | 1,186 | 139 | 0 | 10.5 |
| RMAROUT05 | 596 | 90 | 0 | 13.1 |
| RMAROUT06 | 300 | 59 | 0 | 16.4 |
| RMAROUT07 | 147 | 31 | 0 | 17.4 |
| RMAROUT08 | 76 | 17 | 0 | 18.3 |
| RMAROUT09 | 36 | 6 | 0 | 14.3 |
| RMAROUT10 | 17 | 3 | 0 | 15.0 |
| RMAROUT11 | 8 | 1 | 0 | 11.1 |
| SEX1AGE | 6,776 | 72 | 2 | 1.1 |
| SEX1FOR | 6,776 | 188 | 0 | 2.7 |
| SEX1MTHD1 | 6,785 | 57 | 0 | 0.8 |
| SEX1MTHD2 | 3,328 | 54 | 0 | 1.6 |
| SEX1MTHD3 | 2,673 | 54 | 0 | 2.0 |
| SEX1MTHD4 | 2,645 | 54 | 0 | 2.0 |
| SEX3MO | 6,785 | 71 | 0 | 1.0 |
| SEXEVER | 7,643 | 63 | 0 | 0.8 |
| SEXMAR | 6,785 | 164 | 0 | 2.4 |
| SEXONCE | 6,785 | 2 | 0 | 0.0 |
| SEXP3MO | 6,785 | 63 | 0 | 0.9 |
| SOURCEM1 | 2,489 | 307 | 0 | 11.0 |
| SOURCEM2 | 299 | 53 | 0 | 15.1 |
| URCEM3 | 14 | 10 | 0 | 41.7 |
| URCEM4 | 0 | 10 | 0 | 100.0 |
| TOTINCR | 7,643 | 630 | 0 | 7.6 |
| TUBES | 569 | 7 | 0 | 1.2 |
| VRY1STAG | 6,785 | 30 | 2 | 0.5 |
| VRY1STSX | 6,785 | 30 | 102 | 1.9 |
| WANTP01 | 5,033 | 21 | 0 | 0.4 |
| WANTP02 | 3,766 | 18 | 0 | 0.5 |
| WANTP03 | 2,334 | 21 | 0 | 0.9 |
| WANTP04 | 1,224 | 11 | 0 | 0.9 |
| WANTP05 | 613 | 15 | 0 | 2.4 |
| WANTP06 | 308 | 8 | 0 | 2.5 |
| WANTP07 | 158 | 8 | 0 | 4.8 |
| WANTP08 | 78 | 6 | 0 | 7.1 |
| WANTP5 | 2,823 | 106 | 0 | 3.6 |
| WANTRP01 | 5,033 | 58 | 0 | 1.1 |
| WANTRP02 | 3,766 | 44 | 0 | 1.2 |
| WANTRP03 | 2,334 | 37 | 0 | 1.6 |
| WANTRP04 | 1,224 | 20 | 0 | 1.6 |
| WANTRP05 | 613 | 18 | 0 | 2.9 |
| WANTRP06 | 308 | 5 | 0 | 1.6 |
| WANTRP07 | 158 | 6 | 0 | 3.7 |
| WANTRP08 | 78 | 3 | 0 | 3.7 |
| WANTRP09 | 38 | 1 | 0 | 2.6 |

**Table II. Imputed case counts in the male file**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| ABORTION | 2,190 | 4 | 35 | 1.7 |
| ADDEXP | 4,928 | 17 | 0 | 0.3 |
| AGEBABY1 | 1,731 | 16 | 0 | 0.9 |
| AGEMOMB1 | 4,928 | 64 | 0 | 1.3 |
| B1PREMAR | 1,731 | 4 | 28 | 1.8 |
| CEBOW | 1,731 | 0 | 1 | 0.1 |
| CEBOWC | 1,731 | 0 | 1 | 0.1 |
| CEBOWP | 858 | 4 | 1 | 0.6 |
| COH1DUR | 1,971 | 297 | 1 | 13.1 |
| COHAB1 | 2,115 | 25 | 1 | 1.2 |
| COHOUT | 1,971 | 36 | 1 | 1.8 |
| COHSTAT | 4,928 | 26 | 1 | 0.5 |
| COMPREG | 4,928 | 1 | 23 | 0.5 |
| DATBABY1 | 1,731 | 13 | 0 | 0.7 |
| EDUCAT | 4,928 | 1 | 0 | 0.0 |
| EDUCMOM | 4,928 | 114 | 0 | 2.3 |
| EVHIVTST | 4,928 | 40 | 0 | 0.8 |
| EVRNOPAT | 858 | 4 | 0 | 0.5 |
| FMARNO | 4,928 | 0 | 2 | 0.0 |
| FSEXPAGE | 4,109 | 62 | 21 | 2.0 |
| FSEXRLTN | 4,109 | 84 | 0 | 2.0 |
| HIEDUC | 4,928 | 4 | 0 | 0.1 |
| HISPANIC | 4,928 | 3 | 0 | 0.1 |
| INFEVER | 4,109 | 18 | 0 | 0.4 |
| INSURANC | 4,928 | 80 | 13 | 1.9 |
| INTENT | 4,928 | 3 | 0 | 0.1 |
| LABORFOR | 4,928 | 18 | 0 | 0.4 |
| LIFPRTNR | 4,928 | 162 | 0 | 3.2 |
| LOSSNUM | 2,190 | 4 | 35 | 1.7 |
| LSEXDATE | 4,109 | 74 | 1 | 1.8 |
| LSEXDATE2 | 4,109 | 74 | 1 | 1.8 |
| LSEXPAGE | 4,109 | 46 | 33 | 1.9 |
| LSEXRAGE | 4,109 | 86 | 0 | 2.1 |
| LSEXRLTN | 4,109 | 46 | 4 | 1.2 |
| LSEXUSE1 | 4,109 | 35 | 1 | 0.9 |
| LSEXUSE2 | 1,935 | 42 | 0 | 2.1 |
| LSEXUSE3 | 1,183 | 35 | 0 | 2.9 |
| LSEXUSE4 | 1,111 | 35 | 0 | 3.1 |
| MARBABY1 | 1,731 | 4 | 0 | 0.2 |
| MARDAT01 | 1,754 | 38 | 2 | 2.2 |
| MARDIS01 | 705 | 34 | 3 | 5.0 |
| MAR1DISS | 1,754 | 48 | 3 | 2.8 |
| MAREND1 | 703 | 20 | 1 | 2.9 |
| METH12M1 | 3,578 | 76 | 0 | 2.1 |
| METH12M2 | 1,693 | 80 | 0 | 4.5 |
| METH12M3 | 1,040 | 76 | 0 | 6.8 |
| METH12M4 | 980 | 76 | 0 | 7.2 |
| METH3M1 | 3,071 | 77 | 0 | 2.4 |
| METH3M2 | 1,462 | 81 | 0 | 5.2 |
| METH3M3 | 914 | 77 | 0 | 7.8 |
| METH3M4 | 866 | 77 | 0 | 8.2 |
| NONLIVEB | 4,928 | 1 | 3 | 0.1 |
| NUMP3MOS | 4,109 | 71 | 0 | 1.7 |
| PARAGE14 | 4,928 | 1 | 0 | 0.0 |
| PARTS1YR | 4,928 | 74 | 1 | 1.5 |
| POVERTY | 4,928 | 413 | 0 | 7.7 |
| PREMARW1 | 1,754 | 10 | 0 | 0.6 |
| PUBASSIS | 4,928 | 137 | 0 | 2.7 |
| RELIGION | 4,928 | 20 | 0 | 0.4 |
| SEX12MO | 4,109 | 1 | 69 | 1.7 |
| SEX1MTHD1 | 4,109 | 60 | 0 | 1.4 |
| SEX1MTHD2 | 1,928 | 59 | 0 | 3.0 |
| SEX1MTHD3 | 1,579 | 59 | 0 | 3.6 |
| SEX1MTHD4 | 1,556 | 59 | 0 | 3.7 |
| SEX3MO | 4,109 | 71 | 0 | 1.7 |
| SEXONCE | 4,109 | 1 | 0 | 0.0 |
| SUPP12MO | 629 | 5 | 0 | 0.8 |
| TOTINCR | 4,928 | 413 | 0 | 7.7 |

**Table II. Imputed case counts in the male file—Con.**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| UNINTB5 | 886 | 58 | 1 | 6.2 |
| VRY1STAG | 4,109 | 46 | 2 | 1.2 |
| VRY1STSX | 4,109 | 46 | 129 | 4.1 |
| WANTB1 | 528 | 50 | 0 | 8.7 |
| WANTB2 | 353 | 21 | 0 | 5.6 |
| WANTB3 | 183 | 17 | 0 | 8.5 |
| WANTB4 | 70 | 9 | 0 | 11.4 |
| WANTB5 | 30 | 5 | 0 | 14.3 |
| WANTB6 | 11 | 2 | 0 | 15.4 |
| WANTB8 | 1 | 1 | 0 | 50.0 |

**Table III. Imputed case counts in the pregnancy file**

| Recode variable | Valid N | Regression imputation | Logical imputation | Percent imputed |
|---|---|---|---|---|
| AGECON | 13,593 | 216 | 0 | 1.6 |
| AGEPREG | 13,241 | 175 | 38 | 1.6 |
| BFEEDWKS | 7,998 | 22 | 38 | 0.8 |
| BIRTHORD | 9,148 | 0 | 38 | 0.4 |
| DATECON | 13,593 | 156 | 26 | 1.3 |
| DATEND | 13,241 | 156 | 23 | 1.4 |
| FMARCON5 | 13,593 | 1,009 | 0 | 7.4 |
| FMAROUT5 | 13,240 | 1,092 | 0 | 8.2 |
| LBW1 | 9,148 | 10 | 40 | 0.5 |
| LEARNPRG | 4,052 | 87 | 82 | 4.2 |
| MATERNLV | 3,191 | 18 | 38 | 1.8 |
| OLDWANTP | 13,593 | 79 | 0 | 0.6 |
| OLDWANTR | 13,593 | 192 | 0 | 1.4 |
| OUTCOME | 13,593 | 0 | 38 | 0.3 |
| PAYDELIV | 3,210 | 3 | 38 | 1.3 |
| PMARPREG | 13,241 | 0 | 239 | 1.8 |
| PNCAREWK | 4,052 | 74 | 39 | 2.8 |
| PRGLNGTH | 13,593 | 62 | 0 | 0.5 |
| RMAROUT6 | 13,241 | 1,024 | 0 | 7.7 |
| WANTPART | 13,593 | 108 | 0 | 0.8 |
| WANTRESP | 13,593 | 192 | 0 | 1.4 |

# Appendix III

## Distribution of Female, Male, and Pregnancy Records, by Stratum and Sampling Error Computing Units in the National Survey of Family Growth, Cycle 6

**Table IV. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the combined female and male file**

| Stratum | Statistic | SECU 1 | SECU 2 | Total |
|---|---|---|---|---|
| 1 | Weighted n | 671,524 | 922,508 | 1,594,032 |
| | Sample n | 63 | 75 | 138 |
| | Percent | 42.13 | 57.87 | 100.00 |
| 2 | Weighted n | 3,019,568 | 692,583 | 3,712,151 |
| | Sample n | 280 | 127 | 407 |
| | Percent | 81.34 | 18.66 | 100.00 |
| 3 | Weighted n | 496,943 | 306,710 | 803,653 |
| | Sample n | 47 | 25 | 72 |
| | Percent | 61.84 | 38.16 | 100.00 |
| 4 | Weighted n | 466720 | 212291 | 679011 |
| | Sample n | 39 | 28 | 67 |
| | Percent | 68.74 | 31.26 | 100.00 |
| 5 | Weighted n | 489,087 | 520,163 | 1,009,250 |
| | Sample n | 45 | 50 | 95 |
| | Percent | 48.46 | 51.54 | 100.00 |
| 6 | Weighted n | 535,265 | 457,514 | 992,779 |
| | Sample n | 32 | 27 | 59 |
| | Percent | 53.92 | 46.08 | 100.00 |
| 7 | Weighted n | 462,993 | 443,405 | 906,398 |
| | Sample n | 56 | 44 | 100 |
| | Percent | 51.08 | 48.92 | 100.00 |
| 8 | Weighted n | 369,240 | 788,358 | 1,157,598 |
| | Sample n | 26 | 71 | 97 |
| | Percent | 31.90 | 68.10 | 100.00 |
| 9 | Weighted n | 531,194 | 525,753 | 1,056,947 |
| | Sample n | 95 | 70 | 165 |
| | Percent | 50.26 | 49.74 | 100.00 |
| 10 | Weighted n | 1,349,103 | 882,223 | 2,231,326 |
| | Sample n | 131 | 93 | 224 |
| | Percent | 60.46 | 39.54 | 100.00 |
| 11 | Weighted n | 742,704 | 623,236 | 1,365,940 |
| | Sample n | 90 | 61 | 151 |
| | Percent | 54.37 | 45.63 | 100.00 |
| 12 | Weighted n | 507,099 | 580,743 | 1,087,842 |
| | Sample n | 41 | 46 | 87 |
| | Percent | 46.62 | 53.38 | 100.00 |
| 13 | Weighted n | 227,079 | 318,688 | 545,767 |
| | Sample n | 21 | 34 | 55 |
| | Percent | 41.61 | 58.39 | 100.00 |
| 14 | Weighted n | 1,011,590 | 1,045,454 | 2,057,044 |
| | Sample n | 88 | 79 | 167 |
| | Percent | 49.18 | 50.82 | 100.00 |
| 15 | Weighted n | 831,398 | 776,095 | 1,607,493 |
| | Sample n | 62 | 57 | 119 |
| | Percent | 51.72 | 48.28 | 100.00 |
| 16 | Weighted n | 576,336 | 555,101 | 1,131,437 |
| | Sample n | 59 | 57 | 116 |
| | Percent | 50.94 | 49.06 | 100.00 |
| 17 | Weighted n | 1,287,458 | 1,108,151 | 2,395,609 |
| | Sample n | 134 | 118 | 252 |
| | Percent | 53.74 | 46.26 | 100.00 |
| 18 | Weighted n | 176,286 | 349,629 | 525,915 |
| | Sample n | 25 | 43 | 68 |
| | Percent | 33.52 | 66.48 | 100.00 |
| 19 | Weighted n | 426,169 | 554,926 | 981,095 |
| | Sample n | 35 | 41 | 76 |
| | Percent | 43.44 | 56.56 | 100.00 |
| 20 | Weighted n | 277,366 | 359,145 | 636,511 |
| | Sample n | 28 | 33 | 61 |
| | Percent | 43.58 | 56.42 | 100.00 |

**Table IV. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the combined female and male file—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | SECU Total |
|---|---|---|---|---|
| 21 | Weighted n | 741,249 | 367,624 | 1,108,873 |
| | Sample n | 60 | 38 | 98 |
| | Percent | 66.85 | 33.15 | 100.00 |
| 22 | Weighted n | 1,188,813 | 1,152,949 | 2,341,762 |
| | Sample n | 88 | 87 | 175 |
| | Percent | 50.77 | 49.23 | 100.00 |
| 23 | Weighted n | 696,409 | 867,498 | 1,563,907 |
| | Sample n | 48 | 58 | 106 |
| | Percent | 44.53 | 55.47 | 100.00 |
| 24 | Weighted n | 330,077 | 810,225 | 1,140,302 |
| | Sample n | 24 | 63 | 87 |
| | Percent | 28.95 | 71.05 | 100.00 |
| 25 | Weighted n | 1,089,561 | 1,891,240 | 2,980,801 |
| | Sample n | 97 | 144 | 241 |
| | Percent | 36.55 | 63.45 | 100.00 |
| 26 | Weighted n | 597,452 | 133,032 | 730,484 |
| | Sample n | 29 | 11 | 40 |
| | Percent | 81.79 | 18.21 | 100.00 |
| 27 | Weighted n | 606,471 | 709,396 | 1,315,867 |
| | Sample n | 40 | 91 | 131 |
| | Percent | 46.09 | 53.91 | 100.00 |
| 28 | Weighted n | 655,332 | 827,292 | 1,482,624 |
| | Sample n | 114 | 153 | 267 |
| | Percent | 44.20 | 55.80 | 100.00 |
| 29 | Weighted n | 561,894 | 350,925 | 912,819 |
| | Sample n | 38 | 22 | 60 |
| | Percent | 61.56 | 38.44 | 100.00 |
| 30 | Weighted n | 1,343,591 | 976,877 | 2,320,468 |
| | Sample n | 99 | 77 | 176 |
| | Percent | 57.90 | 42.10 | 100.00 |
| 31 | Weighted n | 969,967 | 969,267 | 1,939,234 |
| | Sample n | 112 | 101 | 213 |
| | Percent | 50.02 | 49.98 | 100.00 |
| 32 | Weighted n | 1,033,768 | 1,109,201 | 2,142,969 |
| | Sample n | 57 | 76 | 133 |
| | Percent | 48.24 | 51.76 | 100.00 |
| 33 | Weighted n | 667,545 | 639,038 | 1,306,583 |
| | Sample n | 90 | 99 | 189 |
| | Percent | 51.09 | 48.91 | 100.00 |
| 34 | Weighted n | 602,221 | 538,225 | 1,140,446 |
| | Sample n | 46 | 44 | 90 |
| | Percent | 52.81 | 47.19 | 100.00 |
| 35 | Weighted n | 916,183 | 787,162 | 1,703,345 |
| | Sample n | 137 | 118 | 255 |
| | Percent | 53.79 | 46.21 | 100.00 |
| 36 | Weighted n | 993,871 | 215,592 | 1,209,463 |
| | Sample n | 67 | 24 | 91 |
| | Percent | 82.17 | 17.83 | 100.00 |
| 37 | Weighted n | 872,294 | 1,043,667 | 1,915,961 |
| | Sample n | 69 | 71 | 140 |
| | Percent | 45.53 | 54.47 | 100.00 |
| 38 | Weighted n | 744,862 | 1,056,404 | 1,801,266 |
| | Sample n | 91 | 68 | 159 |
| | Percent | 41.35 | 58.65 | 100.00 |
| 39 | Weighted n | 575,007 | 435,975 | 1,010,982 |
| | Sample n | 85 | 42 | 127 |
| | Percent | 56.88 | 43.12 | 100.00 |
| 40 | Weighted n | 418,562 | 473,335 | 891,897 |
| | Sample n | 43 | 43 | 86 |
| | Percent | 46.93 | 53.07 | 100.00 |
| 41 | Weighted n | 492,765 | 577,401 | 1,070,166 |
| | Sample n | 59 | 62 | 121 |
| | Percent | 46.05 | 53.95 | 100.00 |
| 42 | Weighted n | 159,296 | 342,615 | 501,911 |
| | Sample n | 12 | 24 | 36 |
| | Percent | 31.74 | 68.26 | 100.00 |

**Table IV. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the combined female and male file—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | SECU Total |
|---|---|---|---|---|
| 43 | Weighted n | 583,233 | 1,107,579 | 1,690,812 |
| | Sample n | 62 | 94 | 156 |
| | Percent | 34.49 | 65.51 | 100.00 |
| 44 | Weighted n | 861,871 | 597,768 | 1,459,639 |
| | Sample n | 149 | 106 | 255 |
| | Percent | 59.05 | 40.95 | 100.00 |
| 45 | Weighted n | 990,736 | 762,649 | 1,753,385 |
| | Sample n | 145 | 100 | 245 |
| | Percent | 56.50 | 43.50 | 100.00 |
| 46 | Weighted n | 744,831 | 1,047,659 | 1,792,490 |
| | Sample n | 113 | 111 | 224 |
| | Percent | 41.55 | 58.45 | 100.00 |
| 47 | Weighted n | 320,256 | 642,505 | 962,761 |
| | Sample n | 28 | 40 | 68 |
| | Percent | 33.26 | 66.74 | 100.00 |
| 48 | Weighted n | 454,169 | 491,608 | 945,777 |
| | Sample n | 57 | 39 | 96 |
| | Percent | 48.02 | 51.98 | 100.00 |
| 49 | Weighted n | 532,638 | 480,304 | 1,012,942 |
| | Sample n | 36 | 39 | 75 |
| | Percent | 52.58 | 47.42 | 100.00 |
| 50 | Weighted n | 811,828 | 564,539 | 1,376,367 |
| | Sample n | 76 | 52 | 128 |
| | Percent | 58.98 | 41.02 | 100.00 |
| 51 | Weighted n | 962,558 | 437,675 | 1,400,233 |
| | Sample n | 62 | 39 | 101 |
| | Percent | 68.74 | 31.26 | 100.00 |
| 52 | Weighted n | 633,713 | 969,912 | 1,603,625 |
| | Sample n | 66 | 86 | 152 |
| | Percent | 39.52 | 60.48 | 100.00 |
| 53 | Weighted n | 764,603 | 855,097 | 1,619,700 |
| | Sample n | 113 | 80 | 193 |
| | Percent | 47.21 | 52.79 | 100.00 |
| 54 | Weighted n | 580,042 | 520,197 | 1,100,239 |
| | Sample n | 88 | 63 | 151 |
| | Percent | 52.72 | 47.28 | 100.00 |
| 55 | Weighted n | 667,986 | 873,011 | 1,540,997 |
| | Sample n | 58 | 100 | 158 |
| | Percent | 43.35 | 56.65 | 100.00 |
| 56 | Weighted n | 1,130,285 | 783,961 | 1,914,246 |
| | Sample n | 133 | 102 | 235 |
| | Percent | 59.05 | 40.95 | 100.00 |
| 57 | Weighted n | 1,576,855 | 1,199,631 | 2,776,486 |
| | Sample n | 137 | 134 | 271 |
| | Percent | 56.79 | 43.21 | 100.00 |
| 58 | Weighted n | 634,245 | 627,104 | 1,261,349 |
| | Sample n | 69 | 72 | 141 |
| | Percent | 50.28 | 49.72 | 100.00 |
| 59 | Weighted n | 813,248 | 702,374 | 1,515,622 |
| | Sample n | 67 | 66 | 133 |
| | Percent | 53.66 | 46.34 | 100.00 |
| 60 | Weighted n | 508,019 | 348,121 | 856,140 |
| | Sample n | 41 | 24 | 65 |
| | Percent | 59.34 | 40.66 | 100.00 |
| 61 | Weighted n | 1,097,764 | 1,108,886 | 2,206,650 |
| | Sample n | 140 | 117 | 257 |
| | Percent | 49.75 | 50.25 | 100.00 |
| 62 | Weighted n | 844,515 | 732,735 | 1,577,250 |
| | Sample n | 77 | 78 | 155 |
| | Percent | 53.54 | 46.46 | 100.00 |
| 63 | Weighted n | 1,293,743 | 1,282,051 | 2,575,794 |
| | Sample n | 123 | 123 | 246 |
| | Percent | 50.23 | 49.77 | 100.00 |
| 64 | Weighted n | 322,331 | 355,637 | 677,968 |
| | Sample n | 27 | 31 | 58 |
| | Percent | 47.54 | 52.46 | 100.00 |

**Table IV. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the combined female and male file—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | Total |
|---|---|---|---|---|
| 65 | Weighted n | 1,314,201 | 1,285,580 | 2,599,781 |
| | Sample n | 203 | 169 | 372 |
| | Percent | 50.55 | 49.45 | 100.00 |
| 66 | Weighted n | 154,945 | 303,011 | 457,956 |
| | Sample n | 18 | 22 | 40 |
| | Percent | 33.83 | 66.17 | 100.00 |
| 67 | Weighted n | 751,028 | 590,712 | 1,341,740 |
| | Sample n | 52 | 46 | 98 |
| | Percent | 55.97 | 44.03 | 100.00 |
| 68 | Weighted n | 327,764 | 244,737 | 572,501 |
| | Sample n | 26 | 21 | 47 |
| | Percent | 57.25 | 42.75 | 100.00 |
| 69 | Weighted n | 849363 | 1,132,988 | 1,982,351 |
| | Sample n | 166 | 178 | 344 |
| | Percent | 42.85 | 57.15 | 100.00 |
| 70 | Weighted n | 658,538 | 691,408 | 1,349,946 |
| | Sample n | 49 | 52 | 101 |
| | Percent | 48.78 | 51.22 | 100.00 |
| 71 | Weighted n | 496,303 | 601,423 | 1,097,726 |
| | Sample n | 72 | 66 | 138 |
| | Percent | 45.21 | 54.79 | 100.00 |
| 72 | Weighted n | 896,616 | 703,720 | 1,600,336 |
| | Sample n | 53 | 60 | 113 |
| | Percent | 56.03 | 43.97 | 100.00 |
| 73 | Weighted n | 856,784 | 879,916 | 1,736,700 |
| | Sample n | 74 | 52 | 126 |
| | Percent | 49.33 | 50.67 | 100.00 |
| 74 | Weighted n | 1,128,429 | 1,358,227 | 2,486,656 |
| | Sample n | 74 | 86 | 160 |
| | Percent | 45.38 | 54.62 | 100.00 |
| 75 | Weighted n | 543,806 | 868,995 | 1,412,801 |
| | Sample n | 83 | 131 | 214 |
| | Percent | 38.49 | 61.51 | 100.00 |
| 76 | Weighted n | 307,946 | 381,518 | 689,464 |
| | Sample n | 39 | 49 | 88 |
| | Percent | 44.66 | 55.34 | 100.00 |
| 77 | Weighted n | 608,677 | 858,176 | 1,466,853 |
| | Sample n | 89 | 106 | 195 |
| | Percent | 41.50 | 58.50 | 100.00 |
| 78 | Weighted n | 2,547,999 | 1,959,068 | 4,507,067 |
| | Sample n | 316 | 264 | 580 |
| | Percent | 56.53 | 43.47 | 100.00 |
| 79 | Weighted n | 481,863 | 531,235 | 1,013,098 |
| | Sample n | 49 | 38 | 87 |
| | Percent | 47.56 | 52.44 | 100.00 |
| 80 | Weighted n | 253,655 | 303,417 | 557,072 |
| | Sample n | 26 | 36 | 62 |
| | Percent | 45.53 | 54.47 | 100.00 |
| 81 | Weighted n | 890,770 | 732,600 | 1,623,370 |
| | Sample n | 76 | 85 | 161 |
| | Percent | 54.87 | 45.13 | 100.00 |
| 82 | Weighted n | 669,916 | 904,813 | 1,574,729 |
| | Sample n | 110 | 100 | 210 |
| | Percent | 42.54 | 57.46 | 100.00 |
| 83 | Weighted n | 157,601 | 272,965 | 430,566 |
| | Sample n | 16 | 27 | 43 |
| | Percent | 36.60 | 63.40 | 100.00 |
| 84 | Weighted n | 736,693 | 839,935 | 1,576,628 |
| | Sample n | 121 | 111 | 232 |
| | Percent | 46.73 | 53.27 | 100.00 |
| Total SECU . . . . . | . . . | . . . | . . . | 122,707,741 |
| Total sample n . . . . . | . . . | . . . | . . . | 12,571 |

. . . Not applicable.

NOTE: n is number.

**Table V. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the pregnancy file, NSFG Cycle 6**

| Stratum | Statistic | SECU 1 | SECU 2 | SECU Total |
|---|---|---|---|---|
| 1 | Weighted n | 335,908 | 1,194,416 | 1,530,324 |
| | Sample n | 48 | 93 | 141 |
| | Percent | 21.95 | 78.05 | 100.00 |
| 2 | Weighted n | 2,155,897 | 838,522 | 2,994,419 |
| | Sample n | 261 | 188 | 449 |
| | Percent | 72.00 | 28.00 | 100.00 |
| 3 | Weighted n | 127,356 | 430,893 | 558,249 |
| | Sample n | 15 | 32 | 47 |
| | Percent | 22.81 | 77.19 | 100.00 |
| 4 | Weighted n | 264,482 | 201,787 | 466,269 |
| | Sample n | 33 | 37 | 70 |
| | Percent | 56.72 | 43.28 | 100.00 |
| 5 | Weighted n | 435,600 | 426,786 | 862,386 |
| | Sample n | 46 | 45 | 91 |
| | Percent | 50.51 | 49.49 | 100.00 |
| 6 | Weighted n | 541,596 | 480,327 | 1,021,923 |
| | Sample n | 41 | 39 | 80 |
| | Percent | 53.00 | 47.00 | 100.00 |
| 7 | Weighted n | 713,984 | 337,650 | 1,051,634 |
| | Sample n | 74 | 40 | 114 |
| | Percent | 67.89 | 32.11 | 100.00 |
| 8 | Weighted n | 207,088 | 287,085 | 494,173 |
| | Sample n | 18 | 37 | 55 |
| | Percent | 41.91 | 58.09 | 100.00 |
| 9 | Weighted n | 572,814 | 475,072 | 1,047,886 |
| | Sample n | 138 | 55 | 193 |
| | Percent | 54.66 | 45.34 | 100.00 |
| 10 | Weighted n | 965,427 | 769,685 | 1,735,112 |
| | Sample n | 139 | 97 | 236 |
| | Percent | 55.64 | 44.36 | 100.00 |
| 11 | Weighted n | 951,734 | 350,263 | 1,301,997 |
| | Sample n | 141 | 80 | 221 |
| | Percent | 73.10 | 26.90 | 100.00 |
| 12 | Weighted n | 476,750 | 511,929 | 988,679 |
| | Sample n | 60 | 54 | 114 |
| | Percent | 48.22 | 51.78 | 100.00 |
| 13 | Weighted n | 88,347 | 263,496 | 351,843 |
| | Sample n | 11 | 36 | 47 |
| | Percent | 25.11 | 74.89 | 100.00 |
| 14 | Weighted n | 910,814 | 1,084,807 | 1,995,621 |
| | Sample n | 116 | 94 | 210 |
| | Percent | 45.64 | 54.36 | 100.00 |
| 15 | Weighted n | 1,152,617 | 1,019,418 | 2,172,035 |
| | Sample n | 90 | 96 | 186 |
| | Percent | 53.07 | 46.93 | 100.00 |
| 16 | Weighted n | 436,044 | 546,520 | 982,564 |
| | Sample n | 51 | 80 | 131 |
| | Percent | 44.38 | 55.62 | 100.00 |
| 17 | Weighted n | 1,098,492 | 1,042,806 | 2,141,298 |
| | Sample n | 146 | 119 | 265 |
| | Percent | 51.30 | 48.70 | 100.00 |
| 18 | Weighted n | 202,321 | 213,388 | 415,709 |
| | Sample n | 47 | 37 | 84 |
| | Percent | 48.67 | 51.33 | 100.00 |
| 19 | Weighted n | 250,722 | 558,546 | 809,268 |
| | Sample n | 23 | 54 | 77 |
| | Percent | 30.98 | 69.02 | 100.00 |
| 20 | Weighted n | 438,486 | 360,849 | 799,335 |
| | Sample n | 46 | 45 | 91 |
| | Percent | 54.86 | 45.14 | 100.00 |
| 21 | Weighted n | 431,655 | 448,916 | 880,571 |
| | Sample n | 43 | 53 | 96 |
| | Percent | 49.02 | 50.98 | 100.00 |
| 22 | Weighted n | 1,352,741 | 922,132 | 2,274,873 |
| | Sample n | 105 | 90 | 195 |
| | Percent | 59.46 | 40.54 | 100.00 |

**Table V. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the pregnancy file, NSFG Cycle 6—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | SECU Total |
|---|---|---|---|---|
| 23 | Weighted n | 685,894 | 1,033,665 | 1,719,559 |
| | Sample n | 56 | 45 | 101 |
| | Percent | 39.89 | 60.11 | 100.00 |
| 24 | Weighted n | 325,124 | 674,948 | 1,000,072 |
| | Sample n | 29 | 48 | 77 |
| | Percent | 32.51 | 67.49 | 100.00 |
| 25 | Weighted n | 936,069 | 1,226,250 | 2,162,319 |
| | Sample n | 113 | 123 | 236 |
| | Percent | 43.29 | 56.71 | 100.00 |
| 26 | Weighted n | 1,991,560 | 174,332 | 2,165,892 |
| | Sample n | 39 | 11 | 50 |
| | Percent | 91.95 | 8.05 | 100.00 |
| 27 | Weighted n | 232,938 | 652,445 | 885,383 |
| | Sample n | 23 | 92 | 115 |
| | Percent | 26.31 | 73.69 | 100.00 |
| 28 | Weighted n | 1,095,420 | 598,117 | 1,693,537 |
| | Sample n | 177 | 167 | 344 |
| | Percent | 64.98 | 35.32 | 100.30 |
| 29 | Weighted n | 301,287 | 341,239 | 642,526 |
| | Sample n | 28 | 27 | 55 |
| | Percent | 46.89 | 53.11 | 100.00 |
| 30 | Weighted n | 1,034,597 | 1,152,505 | 2,187,102 |
| | Sample n | 105 | 111 | 216 |
| | Percent | 47.30 | 52.70 | 100.00 |
| 31 | Weighted n | 1,088,613 | 1,027,288 | 2,115,901 |
| | Sample n | 137 | 109 | 246 |
| | Percent | 51.45 | 48.55 | 100.00 |
| 32 | Weighted n | 1,391,667 | 810,294 | 2,201,961 |
| | Sample n | 64 | 71 | 135 |
| | Percent | 63.20 | 36.80 | 100.00 |
| 33 | Weighted n | 483,099 | 586,482 | 1,069,581 |
| | Sample n | 93 | 73 | 166 |
| | Percent | 45.17 | 54.83 | 100.00 |
| 34 | Weighted n | 487,132 | 495,178 | 982,310 |
| | Sample n | 46 | 56 | 102 |
| | Percent | 49.59 | 50.41 | 100.00 |
| 35 | Weighted n | 637,579 | 830,968 | 1,468,547 |
| | Sample n | 125 | 124 | 249 |
| | Percent | 43.42 | 56.58 | 100.00 |
| 36 | Weighted n | 782,453 | 302,001 | 1,084,454 |
| | Sample n | 68 | 33 | 101 |
| | Percent | 72.15 | 27.85 | 100.00 |
| 37 | Weighted n | 871,661 | 649,849 | 1,521,510 |
| | Sample n | 69 | 60 | 129 |
| | Percent | 57.29 | 42.71 | 100.00 |
| 38 | Weighted n | 958,089 | 700,673 | 1,658,762 |
| | Sample n | 144 | 62 | 206 |
| | Percent | 57.76 | 42.24 | 100.00 |
| 39 | Weighted n | 631,307 | 282,884 | 914,191 |
| | Sample n | 119 | 44 | 163 |
| | Percent | 69.06 | 30.94 | 100.00 |
| 40 | Weighted n | 138,455 | 708,259 | 846,714 |
| | Sample n | 26 | 57 | 83 |
| | Percent | 16.35 | 83.65 | 100.00 |
| 41 | Weighted n | 988,969 | 424,603 | 1,413,572 |
| | Sample n | 86 | 56 | 142 |
| | Percent | 69.96 | 30.04 | 100.00 |
| 42 | Weighted n | 7,396.40 | 396,795 | 404,191.40 |
| | Sample n | 1 | 31 | 32 |
| | Percent | 1.83 | 98.17 | 100.00 |
| 43 | Weighted n | 836,814 | 1,302,174 | 2,138,988 |
| | Sample n | 99 | 105 | 204 |
| | Percent | 39.12 | 60.88 | 100.00 |
| 44 | Weighted n | 879,516 | 299,255 | 1,178,771 |
| | Sample n | 218 | 72 | 290 |
| | Percent | 74.61 | 25.39 | 100.00 |

**Table V. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the pregnancy file, NSFG Cycle 6—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | SECU Total |
|---|---|---|---|---|
| 45 | Weighted n | 1,006,595 | 710,282 | 1,716,877 |
| | Sample n | 184 | 110 | 294 |
| | Percent | 58.63 | 41.37 | 100.00 |
| 46 | Weighted n | 682,761 | 1,250,469 | 1,933,230 |
| | Sample n | 175 | 156 | 331 |
| | Percent | 35.32 | 64.68 | 100.00 |
| 47 | Weighted n | 525,235 | 1,242,201 | 1,767,436 |
| | Sample n | 45 | 64 | 109 |
| | Percent | 29.72 | 70.28 | 100.00 |
| 48 | Weighted n | 581,398 | 519,302 | 1,100,700 |
| | Sample n | 73 | 38 | 111 |
| | Percent | 52.82 | 47.18 | 100.00 |
| 49 | Weighted n | 196,491 | 494,129 | 690,620 |
| | Sample n | 24 | 41 | 65 |
| | Percent | 28.45 | 71.55 | 100.00 |
| 50 | Weighted n | 840,332 | 395,948 | 1,236,280 |
| | Sample n | 97 | 32 | 129 |
| | Percent | 67.97 | 32.03 | 100.00 |
| 51 | Weighted n | 856,934 | 229,050 | 1,085,984 |
| | Sample n | 59 | 31 | 90 |
| | Percent | 78.91 | 21.09 | 100.00 |
| 52 | Weighted n | 537,689 | 910,421 | 1,448,110 |
| | Sample n | 81 | 103 | 184 |
| | Percent | 37.13 | 62.87 | 100.00 |
| 53 | Weighted n | 473,568 | 1,432,568 | 1,906,136 |
| | Sample n | 89 | 144 | 233 |
| | Percent | 24.84 | 75.16 | 100.00 |
| 54 | Weighted n | 643,536 | 932,819 | 1,576,355 |
| | Sample n | 120 | 101 | 221 |
| | Percent | 40.82 | 59.18 | 100.00 |
| 55 | Weighted n | 436,203 | 542,387 | 978,590 |
| | Sample n | 50 | 102 | 152 |
| | Percent | 44.57 | 55.43 | 100.00 |
| 56 | Weighted n | 848,719 | 823,553 | 1,672,272 |
| | Sample n | 105 | 134 | 239 |
| | Percent | 50.75 | 49.25 | 100.00 |
| 57 | Weighted n | 1,523,584 | 1,021,502 | 2,545,086 |
| | Sample n | 147 | 171 | 318 |
| | Percent | 59.86 | 40.14 | 100.00 |
| 58 | Weighted n | 813,488 | 477,789 | 1,291,277 |
| | Sample n | 88 | 69 | 157 |
| | Percent | 63.00 | 37.00 | 100.00 |
| 59 | Weighted n | 672,614 | 603,334 | 1,275,948 |
| | Sample n | 62 | 64 | 126 |
| | Percent | 52.71 | 47.29 | 100.00 |
| 60 | Weighted n | 235,072 | 385,377 | 620,449 |
| | Sample n | 22 | 30 | 52 |
| | Percent | 37.89 | 62.11 | 100.00 |
| 61 | Weighted n | 558,987 | 594,316 | 1,153,303 |
| | Sample n | 89 | 98 | 187 |
| | Percent | 48.47 | 51.53 | 100.00 |
| 62 | Weighted n | 1,060,201 | 808,049 | 1,868,250 |
| | Sample n | 82 | 130 | 212 |
| | Percent | 56.75 | 43.25 | 100.00 |
| 63 | Weighted n | 845,452 | 946,549 | 1,792,001 |
| | Sample n | 79 | 129 | 208 |
| | Percent | 47.18 | 52.82 | 100.00 |
| 64 | Weighted n | 167,157 | 434,743 | 601,900 |
| | Sample n | 19 | 38 | 57 |
| | Percent | 27.77 | 72.23 | 100.00 |
| 65 | Weighted n | 1,505,616 | 1,132,924 | 2,638,540 |
| | Sample n | 274 | 196 | 470 |
| | Percent | 57.06 | 42.94 | 100.00 |
| 66 | Weighted n | 199,483 | 162,752 | 362,235 |
| | Sample n | 33 | 9 | 42 |
| | Percent | 55.07 | 44.93 | 100.00 |

**Table V. Sum of weights and number of cases for Sampling Error Computing Units (SECU) by sampling error stratum for the pregnancy file, NSFG Cycle 6—Con.**

| Stratum | Statistic | SECU 1 | SECU 2 | Total |
|---|---|---|---|---|
| 67 | Weighted n | 594,270 | 435,220 | 1,029,490 |
| | Sample n | 50 | 51 | 101 |
| | Percent | 57.72 | 42.28 | 100.00 |
| 68 | Weighted n | 196,953 | 131,151 | 328,104 |
| | Sample n | 20 | 17 | 37 |
| | Percent | 60.03 | 39.97 | 100.00 |
| 69 | Weighted n | 604,406 | 1,131,221 | 1,735,627 |
| | Sample n | 107 | 204 | 311 |
| | Percent | 34.82 | 65.18 | 100.00 |
| 70 | Weighted n | 352,133 | 603,591 | 955,724 |
| | Sample n | 32 | 63 | 95 |
| | Percent | 36.84 | 63.16 | 100.00 |
| 71 | Weighted n | 537,826 | 392,940 | 930,766 |
| | Sample n | 114 | 41 | 155 |
| | Percent | 57.78 | 42.22 | 100.00 |
| 72 | Weighted n | 652,920 | 762,037 | 1,414,957 |
| | Sample n | 65 | 67 | 132 |
| | Percent | 46.14 | 53.86 | 100.00 |
| 73 | Weighted n | 670,546 | 680,238 | 1,350,784 |
| | Sample n | 50 | 53 | 103 |
| | Percent | 49.64 | 50.36 | 100.00 |
| 74 | Weighted n | 1,079,797 | 1,207,360 | 2,287,157 |
| | Sample n | 65 | 82 | 147 |
| | Percent | 47.21 | 52.79 | 100.00 |
| 75 | Weighted n | 259,188 | 827,277 | 1,086,465 |
| | Sample n | 65 | 129 | 194 |
| | Percent | 23.86 | 76.14 | 100.00 |
| 76 | Weighted n | 186,294 | 228,138 | 414,432 |
| | Sample n | 42 | 22 | 64 |
| | Percent | 44.95 | 55.05 | 100.00 |
| 77 | Weighted n | 514,646 | 462,009 | 976,655 |
| | Sample n | 72 | 109 | 181 |
| | Percent | 52.69 | 47.31 | 100.00 |
| 78 | Weighted n | 1,804,700 | 2,071,197 | 38,75,897 |
| | Sample n | 342 | 306 | 648 |
| | Percent | 46.56 | 53.44 | 100.00 |
| 79 | Weighted n | 189,511 | 441,664 | 631,175 |
| | Sample n | 38 | 29 | 67 |
| | Percent | 30.03 | 69.97 | 100.00 |
| 80 | Weighted n | 226,986 | 222,110 | 449,096 |
| | Sample n | 26 | 39 | 65 |
| | Percent | 50.54 | 49.46 | 100.00 |
| 81 | Weighted n | 1,106,935 | 407,579 | 1,514,514 |
| | Sample n | 107 | 71 | 178 |
| | Percent | 73.09 | 26.91 | 100.00 |
| 82 | Weighted n | 473,905 | 540,277 | 1,014,182 |
| | Sample n | 93 | 81 | 174 |
| | Percent | 46.73 | 53.27 | 100.00 |
| 83 | Weighted n | 96,768 | 295,582 | 392,350 |
| | Sample n | 16 | 41 | 57 |
| | Percent | 24.66 | 75.34 | 100.00 |
| 84 | Weighted n | 479,588 | 953,415 | 1,433,003 |
| | Sample n | 111 | 151 | 262 |
| | Percent | 33.47 | 66.53 | 100.00 |
| Total SECU . . . . . | . . . | . . . | . . . | 111,413,968 |
| Total sample n . . . . . | . . . | . . . | . . . | 13,593 |

. . . Not applicable.

NOTE: n is number.

# Appendix IV

## Variance Estimates Using SAS, SUDAAN, and STATA for a Percentage

Following are programs and output for an analysis of the percentage of females interviewed in NSFG Cycle 6 using the oral contraceptive pill during the month of interview. A cross-tabulation of the use of the oral contraceptive pill by age (in six categories: 15–19, 20–24, 25–29, 30–34, and 40–44 years of age) was generated by SAS 9.1, SUDAAN 8.0.2, STATA 8.0, and WesVar 4.1. The estimates calculated are equivalent across software. Standard errors vary slightly across packages, and design effects vary more substantially. (Output for WesVar is shown only on the Web site to conserve space in this printed report.) SAS data files were converted to STATA 8.0 and SPSS formats using DBMS/COPY 8.0. Variables in upper case are original NSFG Cycle 6 variables or recodes. Variables in lower case represent variables that were recoded as part of the variance estimation program. Library and file names are generic; the user will apply names specific to his or her computing environment. Formatting and library options have been deleted since values will vary across user organizations.

### SAS 9.1

The DATA and SET steps create a dataset for females that contains the variables to be used in the analysis: age categories (''agerx'') and use of contraceptive pill (''pill'').

The PROC SURVEYFREQ produces a cross-tabulation of unweighted and weighted cell counts for the variables (that is, ''agerx'' by ''pill'') specified in the TABLE statement. The WEIGHT statement identifies the weight variable FINALWGT. PROC SURVEYFREQ calculates standard errors appropriate to the complex sample design identified by the STRATUM and CLUSTER statements. The specification of ROW in the TABLE statement limits the cell counts and percentages to the row and DEFF requests calculation of the design effects for the row percentages.

Design effects are greater than 1.0 for all but one of the row proportions due to clustering in the selection and an increase in variance due to weighting. The estimated proportions are equivalent to the other software systems.

### SUDAAN 8.0.2

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA and SET steps used to create a dataset and the variables needed for this analysis (''agerx'' and ''pill''), are identical to those used above in the SAS 9.1 program and are omitted.

The PROC CROSSTAB procedure produces a frequency cross-tabulation of unweighted and weighted cell counts for the analysis variables (agerx by pill) specified in the TABLE statement. The DESIGN used in this computation is specified as WR, with replacement. By specifying the option DEFF in the CROSSTAB statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU_R) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimating the weighted frequency. The specification of NSUM, WSUM, ROWPER, SEROW, and DEFFROW in the PRINT statement limits printed output to row percentages, standard errors of row percentages, and design effects for row percentages.

The estimated percentage of women using a contraceptive pill is identical to those calculated by SAS 9.1 in the six age categories.

### STATA 8.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU_R) variables to be used by STATA 8.0 in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command or running *svyset* again with different settings.

The *generate* and *replace* statements create the recodes ''agerx'' and ''pill.'' The *svytab* command produces a cross-tabulation of ''agerx'' and ''pill'' and provides estimates appropriate to the complex sample design identified by the *svyset* command. The requested estimates and output are limited by specifying *row, deff,* and *se* after the *svyta* command.

Again, the estimated percentages of women using a contraceptive pill in the six age categories are identical to those calculated by SAS 9.1 and SUDAAN 8.0.2.

**SAS 9.1 Program**

```
data NSFG.EX1;
set  NSFG.FEMALES;
if 15 le AGER le 19 then agerx=1;
if 20 le AGER le 24 then agerx=2;
if 25 le AGER le 29 then agerx=3;
if 30 le AGER le 34 then agerx=4;
if 35 le AGER le 39 then agerx=5;
if AGER ge 40 then agerx=6;
if CONSTAT1=6 then pill=1; else pill=2;
run;

proc surveyfreq data=NSFG.EX1;
stratum SEST;
cluster SECU;
weight  FINALWGT;
table agerx*pill / row deff;
run;
```

**SAS 9.1 Output**

The SURVEYFREQ Procedure

Data Summary

| | |
|---|---|
| Number of Strata | 84 |
| Number of Clusters | 168 |
| Number of Observations | 7643 |
| Sum of Weights | 61560714.8 |

Table of agerx by pill

| agerx | pill | Frequency | Weighted Frequency | Std Dev of Wgt Freq | Percent | Std Err of Percent | Design Effect | Row Percent | Std Err of Row Percent |
|---|---|---|---|---|---|---|---|---|---|
| 15-19 | Yes | 187 | 1633986 | 176138 | 2.6543 | 0.2740 | 2.2211 | 16.6155 | 1.4964 |
| | No | 963 | 8200123 | 308550 | 13.3204 | 0.4921 | 1.6029 | 83.3845 | 1.4964 |
| | Total | 1150 | 9834109 | 380244 | 15.9746 | 0.5744 | 1.8784 | 100.000 | |
| 20-24 | Yes | 424 | 3127289 | 338308 | 5.0800 | 0.4776 | 3.6146 | 31.7826 | 1.9966 |
| | No | 939 | 6712331 | 373170 | 10.9036 | 0.4710 | 1.7454 | 68.2174 | 1.9966 |
| | Total | 1363 | 9839620 | 621472 | 15.9836 | 0.7570 | 3.2615 | 100.000 | |
| 25-29 | Yes | 313 | 2366080 | 189219 | 3.8435 | 0.2729 | 1.5400 | 25.5809 | 1.5872 |
| | No | 983 | 6883314 | 377552 | 11.1813 | 0.5279 | 2.1441 | 74.4191 | 1.5872 |
| | Total | 1296 | 9249394 | 467221 | 15.0248 | 0.6057 | 2.1956 | 100.000 | |
| 30-34 | Yes | 275 | 2234545 | 188101 | 3.6298 | 0.2797 | 1.7094 | 21.7527 | 1.4772 |
| | No | 1080 | 8037936 | 396369 | 13.0569 | 0.4906 | 1.6203 | 78.2473 | 1.4772 |
| | Total | 1355 | 10272481 | 477661 | 16.6867 | 0.5571 | 1.7059 | 100.000 | |
| 35-39 | Yes | 170 | 1431768 | 140897 | 2.3258 | 0.2393 | 1.9257 | 13.1922 | 1.2698 |
| | No | 1100 | 9421336 | 427176 | 15.3041 | 0.6189 | 2.2583 | 86.8078 | 1.2698 |
| | Total | 1270 | 10853104 | 441417 | 17.6299 | 0.6615 | 2.3026 | 100.000 | |
| 40-44 | Yes | 98 | 868678 | 98464 | 1.4111 | 0.1540 | 1.3032 | 7.5458 | 0.8347 |
| | No | 1111 | 10643329 | 625810 | 17.2892 | 0.7818 | 3.2666 | 92.4542 | 0.8347 |
| | Total | 1209 | 11512007 | 647860 | 18.7002 | 0.7914 | 3.1481 | 100.000 | |
| Total | Yes | 1467 | 11662345 | 590372 | 18.9445 | 0.6579 | 2.1540 | | |
| | No | 6176 | 49898370 | 1489826 | 81.0555 | 0.6579 | 2.1540 | | |
| | Total | 7643 | 61560715 | 1873490 | 100.000 | | | | |

**SAS 9.1 Output cont.**

```
% Females using oral contraceptive pill                          2

        The SURVEYFREQ Procedure

         Table of AGERX by PILL

                          Row     Std Err of
AGERX     PILL          Percent    Row Percent
_____________________________________________
15-19      Yes          16.6155         1.4964
            No          83.3845         1.4964

         Total          100.000
---------------------------------------------
20-24      Yes          31.7826         1.9966
            No          68.2174         1.9966

         Total          100.000
---------------------------------------------
25-29      Yes          25.5809         1.5872
            No          74.4191         1.5872

         Total          100.000
---------------------------------------------
30-34      Yes          21.7527         1.4772
            No          78.2473         1.4772

         Total          100.000
---------------------------------------------
35-39      Yes          13.1922         1.2698
            No          86.8078         1.2698

         Total          100.000
---------------------------------------------
40-44      Yes           7.5458         0.8347
            No          92.4542         0.8347

         Total          100.000
---------------------------------------------
Total      Yes
            No

         Total
_____________________________________________
```

**SUDAAN Program**

```
(same recode as required in SAS 9.1)

proc sort data=NSFG.EX1;
by SEST SECU_R;
proc crosstab data=NSFG.EX1 design=wr deff;
nest SEST SECU_R;
weight FINALWGT;
subgroup agerx pill;
levels 6 2;
table agerx * pill;
print nsum wsum rowper serow deffrow;
run;
```

## SUDAAN 8.0.2 Output

```
                    S U D A A N
   Software for the Statistical Analysis of Correlated Data
  Copyright      Research Triangle Institute      January 2003
                      Release 8.0.2

Number of observations read    :   7643    Weighted count : 61560715
Denominator degrees of freedom :     84

Variance Estimation Method: Taylor Series (WR)
by: AGERX, using the pill
```

| AGERX | | Using the pill Total | Yes | No |
|---|---|---|---|---|
| Total | Sample Size | 7643.0000 | 1467.0000 | 6176.0000 |
| | Weighted Size | 61560714.7761 | 11662344.8777 | 49898369.8984 |
| | Row Percent | 100.0000 | 18.9445 | 81.0555 |
| | SE Row Percent | 0.0000 | 0.6579 | 0.6579 |
| | DEFF Row Percent #4 | . | 2.1543 | 2.1543 |
| 15-19 | Sample Size | 1150.0000 | 187.0000 | 963.0000 |
| | Weighted Size | 9834108.6926 | 1633985.7873 | 8200122.9053 |
| | Row Percent | 100.0000 | 16.6155 | 83.3845 |
| | SE Row Percent | 0.0000 | 1.4964 | 1.4964 |
| | DEFF Row Percent #4 | . | 1.8587 | 1.8587 |
| 20-24 | Sample Size | 1363.0000 | 424.0000 | 939.0000 |
| | Weighted Size | 9839619.5662 | 3127289.0363 | 6712330.5299 |
| | Row Percent | 100.0000 | 31.7826 | 68.2174 |
| | SE Row Percent | 0.0000 | 1.9966 | 1.9966 |
| | DEFF Row Percent #4 | . | 2.5061 | 2.5061 |
| 25-29 | Sample Size | 1296.0000 | 313.0000 | 983.0000 |
| | Weighted Size | 9249394.2563 | 2366079.9438 | 6883314.3125 |
| | Row Percent | 100.0000 | 25.5809 | 74.4191 |
| | SE Row Percent | 0.0000 | 1.5872 | 1.5872 |
| | DEFF Row Percent #4 | . | 1.7150 | 1.7150 |
| 30-34 | Sample Size | 1355.0000 | 275.0000 | 1080.0000 |
| | Weighted Size | 10272481.3018 | 2234545.0246 | 8037936.2773 |
| | Row Percent | 100.0000 | 21.7527 | 78.2473 |
| | SE Row Percent | 0.0000 | 1.4772 | 1.4772 |
| | DEFF Row Percent #4 | . | 1.7371 | 1.7371 |
| 35-39 | Sample Size | 1270.0000 | 170.0000 | 1100.0000 |
| | Weighted Size | 10853103.9617 | 1431767.5693 | 9421336.3924 |
| | Row Percent | 100.0000 | 13.1922 | 86.8078 |
| | SE Row Percent | 0.0000 | 1.2698 | 1.2698 |
| | DEFF Row Percent #4 | . | 1.7881 | 1.7881 |
| 40-44 | Sample Size | 1209.0000 | 98.0000 | 1111.0000 |
| | Weighted Size | 11512006.9975 | 868677.5165 | 10643329.4810 |
| | Row Percent | 100.0000 | 7.5458 | 92.4542 |
| | SE Row Percent | 0.0000 | 0.8347 | 0.8347 |
| | DEFF Row Percent #4 | . | 1.2075 | 1.2075 |

**STATA 8.0 Program**

```
use "EX1.DTA"

svyset [pweight=FINALWGT], strata(SEST) psu(SECU_R)

generate agerx=1 if AGER <=19
replace  agerx=2 if AGER >=20 & AGER <=24
replace  agerx=3 if AGER >=25 & AGER <=29
replace  agerx=4 if AGER >=30 & AGER <=34
replace  agerx=5 if AGER >=35 & AGER <=39
replace  agerx=6 if AGER >=40

generate pill=2
replace pill=1 if CONSTAT1==6

svytab agerx pill, row se deff percent
```

**STATA 8.0 Output**

```
pweight:  finalwgt                          Number of obs      =      7643
Strata:   sest                              Number of strata   =        84
PSU:      secu_r                            Number of PSUs     =       168
                                            Population size    =  61560715


          |       Using the pill
          |
    agerx |     Yes        No    Total
----------+-------------------------
    15-19 |   16.62     83.38      100
          |  (1.496)   (1.496)
          |   66.23     14.82
          |
    20-24 |   31.78     68.22      100
          |  (1.997)   (1.997)
          |   63.18     31.36
          |
    25-29 |   25.58     74.42      100
          |  (1.587)   (1.587)
          |   52.09     19.39
          |
    30-34 |   21.75     78.25      100
          |  (1.477)   (1.477)
          |   47.67     14.69
          |
    35-39 |   13.19     86.81      100
          |   (1.27)    (1.27)
          |   54.24     9.506
          |
    40-44 |   7.546     92.45      100
          |  (.8347)   (.8347)
          |   38.28     3.724
          |
    Total |   18.94     81.06      100
          |  (.6579)   (.6579)
          |   2.154     2.154

  Key:  row percentages
        (standard errors of row percentages)
        deff for variances of row percentages

  Pearson:
    Uncorrected   chi2(5)          =  324.8924
    Design-based  F(4.63, 388.69)  =   36.6663     P = 0.0000

  Mean generalized deff            =    2.0255
  CV of generalized deffs          =    0.5664
```

# Appendix V

## Variance Estimates Using SAS, SUDAAN, and STATA, for a Mean

Following are the programs and output for an analysis of the mean number of children born to women 20–44 years of age interviewed in Cycle 6 of the NSFG, by race and Hispanic origin and age. Means were generated by SAS 9.1, SUDAAN 8.0.2, STATA 8.0, and WesVar 4.1. The estimates calculated are equivalent across software. However, due to specific methods used in calculations, standard errors vary slightly across packages, and design effects vary more substantially. (Output for WesVar is shown only on the Web site to conserve space in this printed report.) SAS data files were converted to STATA 8.0 and SPSS formats (for use in WesVar 4.1) using DBMS/COPY 8.0. Variables in upper case are original NSFG Cycle 6 variables or recodes. Variables in lower case represent variables that were recoded as part of the variance estimation program. Library and file names are generic and it is assumed the user will apply names specific to his or her computing environment. Formatting and library options have been deleted; values will vary across user organizations.

### SAS 9.1

The DATA and SET steps create a dataset that contains the variables for females to be used in the analysis. The PROC SURVEYMEANS produces a table of weighted means for the variable specified in the VAR statement. The WEIGHT statement identifies the weight variable (FINALWGT) to be used in estimating the means. PROC SURVEYMEANS calculates standard errors appropriate to the complex sample design specified in the STRATUM and CLUSTER statements. The DEFF option is not available with PROC SURVEYMEANS.

The estimated weighted means are equivalent to the other software systems.

**SAS 9.1 Program**

```
data NSFG.EX6;
set  NSFG.FEMALES;
if AGER lt 20 then delete;
if 20 le AGER le 24 then agerx=1;
if 25 le AGER le 34 then agerx=2;
if AGER ge 35       then agerx=3;
if HISPRACE=1 and agerx=1 then group=1;
if HISPRACE=1 and agerx=2 then group=2;
if HISPRACE=1 and agerx=3 then group=3;
if HISPRACE=2 and agerx=1 then group=4;
if HISPRACE=2 and agerx=2 then group=5;
if HISPRACE=2 and agerx=3 then group=6;
if HISPRACE=3 and agerx=1 then group=7;
if HISPRACE=3 and agerx=2 then group=8;
if HISPRACE=3 and agerx=3 then group=9;
if HISPRACE=4 and agerx=1 then group=10;
if HISPRACE=4 and agerx=2 then group=11;
if HISPRACE=4 and agerx=3 then group=12;
run;

proc sort data=NSFG.EX6;
by group;
proc surveymeans data=NSFG.EX6;
stratum SEST;
cluster SECU_R;
weight  FINALWGT;
var PARITY;
by group;
run;
```

### SUDAAN 8.0.2

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA and SET statements used to create a dataset and the variables needed for this analysis (“agerx” and “group”) are identical to those used above in the SAS 9.1 program and are omitted for this program. The PROC DESCRIPT procedure produces descriptive statistics for the variable specified in the VAR statement (PARITY). The DESIGN used in this analysis is WR, with replacement. By specifying DEFF in the DESCRIPT statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU_R) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimating the weighted means. SUBGROUP identifies subpopulations for which estimates are requested.

The estimated mean number of children born to women 20–44 years of age by race and Hispanic origin and age is identical to those calculated by SAS 9.1.

### STATA 8.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU_R) variables to be used by STATA 8.0 in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command or running *svyset* again with different settings.

The *generate* and *replace* statements create the recodes “agerx” and “group.” The *svymean* command produces estimated weighted means for each of the levels of the *by* variable “group.” The estimates provided are appropriate to the complex sample design identified by the *svyset* command. Design effect calculations are requested by entering *deff* after the *svymean* command.

Again, the estimated mean number of children born by race and Hispanic origin and age is identical to those calculated by SAS 9.1 and SUDAAN 8.0.2.

## SAS 9.1 Output

Mean number of children born to women 20-44 years of age

The SURVEYMEANS Procedure

Data Summary

| | |
|---|---|
| Number of Strata | 84 |
| Number of Clusters | 168 |
| Number of Observations | 6493 |
| Sum of Weights | 51726606.1 |

Statistics

| Variable | Label | N | Mean | Std Error of Mean |
|---|---|---|---|---|
| PARITY | TOTAL NUMBER OF LIVE BIRTHS | 6493 | 1.502092 | 0.038181 |

Statistics

| Variable | 95% CL for Mean | |
|---|---|---|
| PARITY | 1.42616568 | 1.57801861 |

Domain Analysis: group

| group | Variable | Label | N | Mean |
|---|---|---|---|---|
| Hispanic 20-24 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 293 | 0.917116 |
| Hispanic 25-34 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 605 | 1.840386 |
| Hispanic 35-44 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 460 | 2.494708 |
| Non-Hisp White 20-24 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 775 | 0.338462 |
| Non-Hisp White 25-34 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 1331 | 1.229966 |
| Non-Hisp White 35-44 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 1420 | 1.928667 |

Domain Analysis: group

| group | Variable | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|
| Hispanic 20-24 | PARITY | 0.089134 | 0.74052617 | 1.09370619 |
| Hispanic 25-34 | PARITY | 0.095041 | 1.65201805 | 2.02875331 |
| Hispanic 35-44 | PARITY | 0.088816 | 2.31873168 | 2.67068473 |
| Non-Hisp White 20-24 | PARITY | 0.029326 | 0.28016468 | 0.39676002 |
| Non-Hisp White 25-34 | PARITY | 0.048924 | 1.13269100 | 1.32724049 |
| Non-Hisp White 35-44 | PARITY | 0.062855 | 1.80367322 | 2.05366074 |

The SURVEYMEANS Procedure

Domain Analysis: group

| group | Variable | Label | N | Mean |
|---|---|---|---|---|
| Non-Hisp Black 20-24 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 218 | 0.795388 |
| Non-Hisp Black 25-34 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 568 | 1.820761 |
| Non-Hisp Black 35-44 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 502 | 2.137208 |
| Non-Hisp Other 20-24 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 77 | 0.584299 |
| Non-Hisp Other 25-34 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 147 | 1.083117 |
| Non-Hisp Other 35-44 | PARITY | TOTAL NUMBER OF LIVE BIRTHS | 97 | 1.835584 |

Domain Analysis: group

| group | Variable | Std Error of Mean | 95% CL for Mean | |
|---|---|---|---|---|
| Non-Hisp Black 20-24 | PARITY | 0.098256 | 0.60072506 | 0.99005081 |
| Non-Hisp Black 25-34 | PARITY | 0.072588 | 1.67684760 | 1.96467420 |
| Non-Hisp Black 35-44 | PARITY | 0.090084 | 1.95858741 | 2.31582897 |
| Non-Hisp Other 20-24 | PARITY | 0.181311 | 0.22548892 | 0.94310855 |
| Non-Hisp Other 25-34 | PARITY | 0.124411 | 0.83668391 | 1.32955101 |
| Non-Hisp Other 35-44 | PARITY | 0.269601 | 1.30169962 | 2.36946800 |

## SUDAAN 8.0.2 Program

```
(same recode as required in SAS9)

proc sort data=NSFG.EX6;
by SEST SECU_R;
proc descript data=NSFG.EX6 design=wr deff;
nest SEST SECU_R;
weight  FINALWGT;
var PARITY;
subgroup group;
levels 12;
run;
```

## SUDAAN 8.0.2 Output

Mean number of children born to women 20-44 by hisprace and age

S U D A A N
Software for the Statistical Analysis of Correlated Data
Copyright Research Triangle Institute January 2003
Release 8.0.2

Number of observations read : 6493 Weighted count : 51726606
Denominator degrees of freedom : 84

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

| Variable | | GROUP Total | Hispanic 20-24 | Hispanic 25-34 |
|---|---|---|---|---|
| TOTAL NUMBER OF LIVE BIRTHS | Sample Size | 6493 | 293 | 605 |
| | Weighted Size | 51726606.08 | 1632242.37 | 3248970.16 |
| | Total | 77698128.71 | 1496955.88 | 5979358.16 |
| | Mean | 1.50 | 0.92 | 1.84 |
| | SE Mean | 0.04 | 0.09 | 0.10 |
| | DEFF Mean #4 | 4.71 | 2.11 | 2.90 |
| | DEFF Total #4 | 8.78 | 4.98 | 8.22 |

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

| Variable | | GROUP Hispanic 35-44 | Non-Hisp White 20-24 | Non-Hisp White 25-34 |
|---|---|---|---|---|
| TOTAL NUMBER OF LIVE BIRTHS | Sample Size | 460 | 775 | 1331 |
| | Weighted Size | 2704983.78 | 6071689.25 | 12304528.28 |
| | Total | 6748145.22 | 2055038.20 | 15134148.32 |
| | Mean | 2.49 | 0.34 | 1.23 |
| | SE Mean | 0.09 | 0.03 | 0.05 |
| | DEFF Mean #4 | 1.31 | 1.36 | 2.13 |
| | DEFF Total #4 | 5.62 | 1.56 | 4.41 |

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

| Variable | | GROUP Non-Hisp White 35-44 | Non-Hisp Black 20-24 | Non-Hisp Black 25-34 |
|---|---|---|---|---|
| TOTAL NUMBER OF LIVE BIRTHS | Sample Size | 1420 | 218 | 568 |
| | Weighted Size | 15788624.63 | 1457688.28 | 2694783.22 |
| | Total | 30450999.01 | 1159427.67 | 4906555.91 |
| | Mean | 1.93 | 0.80 | 1.82 |
| | SE Mean | 0.06 | 0.10 | 0.07 |
| | DEFF Mean #4 | 3.10 | 1.60 | 1.24 |
| | DEFF Total #4 | 12.95 | 1.69 | 4.35 |

**SUDAAN 8.0.2 Output cont.**

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

| Variable | | GROUP<br>Non-Hisp<br>Black 35-44 | Non-Hisp<br>Other 20-24 | Non-Hisp<br>Other 25-34 |
|---|---|---|---|---|
| TOTAL NUMBER OF LIVE BIRTHS | Sample Size | 502 | 77 | 147 |
| | Weighted Size | 2935527.92 | 677999.66 | 1273593.90 |
| | Total | 6273834.31 | 396154.34 | 1379451.79 |
| | Mean | 2.14 | 0.58 | 1.08 |
| | SE Mean | 0.09 | 0.18 | 0.12 |
| | DEFF Mean #4 | 1.61 | 2.77 | 1.63 |
| | DEFF Total #4 | 6.92 | 4.04 | 3.21 |

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

| Variable | | GROUP<br>Non-Hisp<br>Other 35-44 |
|---|---|---|
| TOTAL NUMBER OF LIVE BIRTHS | Sample Size | 97 |
| | Weighted Size | 935974.64 |
| | Total | 1718059.89 |
| | Mean | 1.84 |
| | SE Mean | 0.27 |
| | DEFF Mean #4 | 3.52 |
| | DEFF Total #4 | 8.83 |

**STATA 8.0 Program**

```
use "females.dta"

drop if ager < 20

svyset [pweight=FINALWGT], strata(SEST) psu(SECU_R)

generate agerx=1 if AGER >=20 & AGER <=24
replace  agerx=2 if AGER >=25 & AGER <=34
replace  agerx=3 if AGER >=35

generate group=1  if HISPRACE ==1 & agerx=1
replace  group=2  if HISPRACE ==1 & agerx=2
replace  group=3  if HISPRACE ==1 & agerx=3
replace  group=4  if HISPRACE ==2 & agerx=1
replace  group=5  if HISPRACE ==2 & agerx=2
replace  group=6  if HISPRACE ==2 & agerx=3
replace  group=7  if HISPRACE ==3 & agerx=1
replace  group=8  if HISPRACE ==3 & agerx=2
replace  group=9  if HISPRACE ==3 & agerx=3
replace  group=10 if HISPRACE ==4 & agerx=1
replace  group=11 if HISPRACE ==4 & agerx=2
replace  group=12 if HISPRACE ==4 & agerx=3

svymean PARITY, by(group) deff
```

**STAT 8.0 Output**

```
. svymean parity, by(group) deff

Survey mean estimation

pweight:  finalwgt                          Number of obs     =      6493
Strata:   sest                              Number of strata  =        84
PSU:      secu_r                            Number of PSUs    =       168
                                            Population size   =  51726606
```

| Mean | Subpop. | Estimate | Std. Err. | Deff |
|---|---|---|---|---|
| parity | | | | |
| Hispanic | 20-24 | .9171162 | .0891337 | 1.478611 |
| Hispanic | 25-34 | 1.840386 | .0950407 | 1.956508 |
| Hispanic | 35-44 | 2.494708 | .0888155 | .9661928 |
| White | 20-24 | .3384623 | .0293258 | 1.339018 |
| White | 25-34 | 1.229966 | .0489244 | 2.468895 |
| White | 35-44 | 1.928667 | .0628548 | 4.335046 |
| Black | 20-24 | .7953879 | .098256 | 1.351903 |
| Black | 25-34 | 1.820761 | .0725883 | .7424528 |
| Black | 35-44 | 2.137208 | .0900844 | 1.187491 |
| Other | 20-24 | .5842987 | .1813113 | 3.105673 |
| Other | 25-34 | 1.083117 | .1244106 | 1.789126 |
| Other | 35-44 | 1.835584 | .2696014 | 4.31016 |

# Appendix VI

## Variance Estimates Using SAS, SUDAAN, and STATA for a Regression

Following are the programs and output for an analysis of the relationship among the number of children born to women 20–44 years of age interviewed in Cycle 6 of the NSFG, race and Hispanic origin, and education. Coefficients were generated by SAS 9.1, SUDAAN 8.0.2, STATA 8.0, and WesVar 4.1. The estimates calculated are equivalent across software. However, due to specific methods used in calculations, standard errors vary slightly across packages, and design effects vary more substantially. (Output for WesVar is shown only on the web site to conserve space in this printed report.)

SAS data files were converted to STATA 8.0 and SPSS formats using DBMS/COPY 8.0. Variables in upper case are original NSFG Cycle 6 variables or recodes. Variables in lower case represent variables that were recoded as part of the variance estimation program. Library and file names are generic and it is assumed the user will apply names specific to his or her computing environment. Formatting and library options have been deleted; preferences will vary across user organizations.

### SAS 9.1

The DATA and SET steps create a dataset which contains the variables for females to be used in the analysis. The PROC SURVEYREG estimates the relationship between a continuous variable (PARITY) and a set of predictors (AGER, ''hieducx'', and ''black'') specified by the MODEL statement. The WEIGHT statement identifies the weight variable (FINALWGT) to be used in estimating the means. PROC SURVEYREG calculates standard errors appropriate to the complex sample design specified in the STRATUM and CLUSTER statements. The DEFF option requests the calculation of design effects.

**SAS 9.1 Program**

```
data NSFG.EX7;
set NSFG.FEMALES;
if AGER lt 20 then delete;
if HISPRACE=3 then black=1;              *black
if HISPRACE in (1 2 4) then black=0;     *non-black
if HIEDUC le 9 then hieducx=0;           *high school or less;
else if HIEDUC gt 9 then hieducx=1;      *more than high school;
run;

proc surveyreg data=NSFG.EX7;
stratum SEST;
cluster SECU_R;
weight  FINALWGT;
model PARITY= AGER hieducx black / deff;
run;
```

**SAS 9.1 Output**

```
Female Parity regressed on race and ethnicity, age, and education

                    The SURVEYREG Procedure

          Regression Analysis for Dependent Variable PARITY

                         Data Summary

          Number of Observations              6493
          Sum of Weights                  51726606
          Weighted Mean of PARITY          1.50209
          Weighted Sum of PARITY          77698129

                         Design Summary

             Number of Strata                 84
             Number of Clusters              168

                         Fit Statistics

                R-square              0.2282
                Root MSE              1.2460
                Denominator DF            84

                     Tests of Model Effects

          Effect        Num DF    F Value    Pr > F

          Model              3     372.67    <.0001
          Intercept          1      35.60    <.0001
          AGER               1     587.26    <.0001
          hieducx            1     386.16    <.0001
          black              1      15.17    0.0002

     NOTE: The denominator degrees of freedom for the F tests is 84.

                  Estimated Regression Coefficients

                                Standard                                Design
Parameter       Estimate          Error    t Value     Pr > |t|        Effect

Intercept     -0.5559592     0.09317570      -5.97       <.0001          1.58
AGER           0.0760041     0.00313634      24.23       <.0001          0.00
hieducx       -0.7451044     0.03791680     -19.65       <.0001          1.45
black          0.2244115     0.05761279       3.90       0.0002          1.63

     NOTE: The denominator degrees of freedom for the t tests is 84.
```

The estimated regression coefficients are equivalent to the other software systems.

### SUDAAN 8.0.2

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA and SET statements used to create a dataset and the variables needed for this analysis are identical to those used above in the SAS 9.1 program and are omitted for this program.

The PROC REGRESS estimates the relationship between a continuous variable (PARITY) and a set of predictors (AGER, ''hieducx'', and ''black'') specified by the MODEL statement. The DESIGN used in this

analysis is WR, with replacement. By specifying DEFT4 in the REGRESS statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU_R) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimated the weighted means.

The estimated coefficients calculated by SUDAAN 8.0.2 are identical to those from SAS 9.1.

**SUDAAN 8.0.2 Program**

```
(same recode as required in sas9)

proc sort data=NSFG.EX7;
by SEST SECU_R;
proc regress data=NSFG.EX7 design=wr deft4;
nest SEST SECU_R;
weight  FINALWGT;
model PARITY=AGER hieducx black;
run;
```

## STATA 8.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU_R) variables to be used by STATA 8.0 in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command or running *svyset* again with different settings.

The *generate* and *replace* statements create the recodes ‘‘hieducx’’ and ‘‘black.’’ The *svyreg* command models the relationship between PARITY and a set of predictors (AGER, ‘‘hieducx,’’ and ‘‘black’’). The estimates provided are appropriate to the complex sample design identified by the *svyset* command. Design effect calculations are requested by entering *deff* after the *svyreg* command.

The estimated coefficients as calculated by STATA 8.0 are identical to those calculated by SAS 9.1 and SUDAAN 8.0.2.

**SUDAAN 8.0.2 Output**

```
Female Parity regressed on race and ethnicity, age, and education
                              S U D A A N
      Software for the Statistical Analysis of Correlated Data
    Copyright      Research Triangle Institute      January 2003
                          Release 8.0.2

Number of observations read        :   6493    Weighted count: 51726606
Observations used in the analysis  :   6493    Weighted count: 51726606
Denominator degrees of freedom     :     84

Maximum number of estimable parameters for the model is  4

File NSFG.EX7 contains  168 Clusters
 168 clusters were used to fit the model
Maximum cluster size is 155 records
Minimum cluster size is   3 records

Weighted mean response is 1.502092

Multiple R-Square for the dependent variable PARITY: 0.228243

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable PARITY: TOTAL NUMBER OF LIVE BIRTHS
```

| Independent Variables and Effects | Beta Coeff. | DEFF Beta #4 | SE Beta | T-Test B=0 | P-value T-Test B=0 |
|---|---|---|---|---|---|
| Intercept | -0.56 | 1.58 | 0.09 | -5.97 | 0.0000 |
| AGE AT INTERVIEW | 0.08 | 2.17 | 0.00 | 24.24 | 0.0000 |
| HIEDUCX | -0.75 | 1.45 | 0.04 | -19.66 | 0.0000 |
| BLACK | 0.22 | 1.63 | 0.06 | 3.90 | 0.0002 |

## SUDAAN 8.0.2 Output

```
Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Identity
Response variable PARITY: TOTAL NUMBER OF LIVE BIRTHS
-------------------------------------------------------

Contrast                Degrees
                        of                    P-value
                        Freedom       Wald F  Wald F
-------------------------------------------------------
OVERALL MODEL                 4       780.81    0.0000
MODEL MINUS
  INTERCEPT                   3       372.84    0.0000
INTERCEPT                     1        35.62    0.0000
AGER                          1       587.53    0.0000
HIEDUCX                       1       386.34    0.0000
BLACK                         1        15.18    0.0002
-------------------------------------------------------
```

## STATA 8.0 Program

```
use "females.dta"

svyset [pweight=FINALWGT], strata(SEST) psu(SECU_R)

drop if AGER <20
generate hieducx=0 if HIEDUC <=9
replace  hieducx=1 if HIEDUC >9

generate black=0
replace  black=1 if HISPRACE==3

svyreg PARITY AGER hieducx black, deff
```

## STATA 8.0 Output

```
. svyreg parity ager hieducx black, deff

Survey linear regression

pweight:  finalwgt                          Number of obs     =      6493
Strata:   sest                              Number of strata  =        84
PSU:      secu_r                            Number of PSUs    =       168
                                            Population size   =  51726606
                                            F(   3,      82)  =    363.97
                                            Prob > F          =    0.0000
                                            R-squared         =    0.2282

------------------------------------------------------------------------------
      parity |      Coef.   Std. Err.      Deff
-------------+----------------------------------------------------------------
        ager |   .0760041   .0031356   2.428935
     hieducx |  -.7451044    .037908   1.322322
       black |   .2244115   .0575995   1.274843
       _cons |  -.5559592   .0931542   1.994361
------------------------------------------------------------------------------
```

# Appendix VII

## Variance Estimates in SAS, SUDAAN, and STATA for a Logistic Regression

Following are the programs and output for an analysis of the probability of strongly agreeing that "a young couple should not live together unless they are married." An indicator variable of strong agreement was regressed on age, gender, race and Hispanic origin, and education. Regression coefficients and odds ratios were generated by SAS 9.1, SUDAAN 8.0.2, STATA 8.0, and WesVar 4.1. The estimates calculated are equivalent across software. However, due to specific methods used in calculations, standard errors vary slightly across packages, and design effects vary more substantially. (Output for WesVar is shown only on the Web site to conserve space in this printed report.)

SAS data files were converted to STATA 8.0 and SPSS formats (for use in WesVar 4.1) using DBMS/COPY 8.0. Variables in upper case are original NSFG Cycle 6 variables or recodes. Variables in lower case represent variables that were recoded as part of the variance estimation program. Library and file names are generic and it is assumed the user will apply names specific to his or her computing environment. Formatting and library options have been deleted; values will vary across user organizations.

### SAS 9.1

The DATA, SET, and MERGE steps create a dataset that contains the variables and recodes ("okcohabx," "black," and "hieducx") for males and females to be used in the analysis. The PROC SURVEYLOGISTIC models the relationship between a dichotomous variable ("okcohabx") and a set of predictors (AGER, "hieducx," "black," and FEMALE) specified in the MODEL statement. The WEIGHT statement identifies the weight variable (FINALWGT) to be used in estimating the model. PROC SURVEYLOGISTIC calculates standard errors appropriate to the complex sample design specified in the STRATUM and CLUSTER statements. The DEFF option, which requests calculation of design effects, is not available with PROC SURVEYLOGISTIC.

The estimated coefficients and odds ratios are equivalent to the other software systems.

### SUDAAN 8.0.2

A SAS-callable version of SUDAAN 8.0.2 was used to calculate the estimates for this example. The DATA, SET, and MERGE statements used to create a dataset and the variables needed for this analysis are identical to those used previously in the SAS 9.1 program and are omitted for this program.

The PROC RLOGIST estimates the relationship between a dichotomous variable ("okcohabx") and a set of predictors (AGER, "hieducx," "black," and FEMALE) specified in the MODEL statement. The DESIGN used in this analysis is WR, with replacement. By specifying DEFT4 in the RLOGIST statement, design effects will be calculated. The NEST statement specifies the strata (SEST) and cluster (SECU) variables for calculating standard errors appropriate to the complex sample design. The WEIGHT statement identifies FINALWGT for estimation.

The estimated coefficients and odds ratios calculated by SUDAAN 8.0.2 are identical to those from SAS 9.1.

**SAS 9.1 Program**

```
data mlivtog (keep=CASEID AGER FEMALE HISPRACE HIEDUC SECU SEST FINALWGT OKCOHAB);
set NSFG.MALES;
data flivtog (keep= CASEID AGER FEMALE HISPRACE HIEDUC SECU_R SECU SEST FINALWGT
OKCOHAB);
set NSFG.FEMALES;
SECU=SECU_R;
run;

proc sort data=mlivtog;   by CASEID;
proc sort data=flivtog;   by CASEID;
data NSFG.MF_LIVTOG;      merge mlivtog flivtog;  by CASEID;  run;

data NSFG.EX8;
set NSFG.MF_LIVTOG;
if OKCOHAB=1 then okcohabx=1;   else okcohabx=5;
if HISPRACE in (1 2 4) then black=0;                *Non-black
if HISPRACE=3 then black=1;                         *Black
if HIEDUC le 9 then hieducx=0;                      *high school or less;
else if HIEDUC gt 9 then hieducx=1;                 *more than high school;
run;

proc surveylogistic data=NSFG.EX8;
stratum SEST;
cluster SECU;
weight FINALWGT;
model okcohabx= AGER hieducx black FEMALE;
run;
```

## STATA 8.0

The *use* statement specifies the dataset to be used. The *svyset* command specifies the weight (FINALWGT), strata (SEST), and cluster (SECU) variables to be used by STATA 8.0 in estimation. These settings are saved for the current session, but can be cleared by entering the *clear* command or running *svyset* again with different settings.

The *generate* and *replace* statements create recodes: ''okcohabx,'' ''hieducx,'' and ''black.'' The *svylogit* the relationship between a dichotomous variable (''okcohabx'') and a set of predictors (AGER, ''hieducx,'' ''black,'' and FEMALE) specified in the *svylogit* statement. The estimates provided are appropriate to the complex sample design identified by the *svyset* command. Design effect calculations are requested by entering *deff* after the *svylogit* command.

The estimated coefficients as calculated by STATA 8.0 are identical to those calculated by SAS 9.1 and SUDAAN 8.0.2.

**SAS 9.1 Output**

Probability of strongly agreeing that a young couple should not live together unless married

The SURVEYLOGISTIC Procedure

Testing Global Null Hypothesis: BETA=0

| Test | Chi-Square | DF | Pr > ChiSq |
|---|---|---|---|
| Likelihood Ratio | 433975.770 | 4 | <.0001 |
| Score | 435582.565 | 4 | <.0001 |
| Wald | 24.7892 | 4 | <.0001 |

Analysis of Maximum Likelihood Estimates

| Parameter | DF | Estimate | Standard Error | Wald Chi-Square | Pr > ChiSq |
|---|---|---|---|---|---|
| Intercept | 1 | -2.7277 | 0.2099 | 168.9313 | <.0001 |
| AGER | 1 | 0.00723 | 0.00703 | 1.0579 | 0.3037 |
| hieducx | 1 | 0.3204 | 0.1208 | 7.0301 | 0.0080 |
| black | 1 | 0.2710 | 0.1063 | 6.4945 | 0.0108 |
| female | 1 | 0.0626 | 0.1033 | 0.3675 | 0.5443 |

Odds Ratio Estimates

| Effect | Point Estimate | 95% Wald Confidence Limits | |
|---|---|---|---|
| AGER | 1.007 | 0.993 | 1.021 |
| hieducx | 1.378 | 1.087 | 1.746 |
| black | 1.311 | 1.065 | 1.615 |
| female | 1.065 | 0.870 | 1.304 |

Association of Predicted Probabilities and Observed Responses

| | | | |
|---|---|---|---|
| Percent Concordant | 54.7 | Somers' D | 0.133 |
| Percent Discordant | 41.4 | Gamma | 0.139 |
| Percent Tied | 3.9 | Tau-a | 0.021 |
| Pairs | 12285204 | c | 0.567 |

**SUDAAN 8.0.2 Program**

```
(same recode as required in SAS9)

proc sort data=NSFG.EX8;
by SEST SECU;
proc rlogist data=NSFG.EX8 design=wr deft4;
weight  FINALWGT;
nest SEST SECU;
model okcohabx = AGER hieducx black FEMALE;
run;
```

## SUDAAN 8.0.2 Output

```
Agreement couples should not live together unless married regressed on age race gendereducation

                                S U D A A N
          Software for the Statistical Analysis of Correlated Data
        Copyright      Research Triangle Institute      January 2003
                              Release 8.0.2

Number of zero responses      : 11503
Number of non-zero responses :  1068

Independence parameters have converged in 5 iterations

Number of observations read       :  12571    Weighted count:122707736
Observations used in the analysis :  12571    Weighted count:122707736
Denominator degrees of freedom    :     84

Maximum number of estimable parameters for the model is  5

File NSFG.EX8X contains  168 Clusters
 168 clusters were used to fit the model
Maximum cluster size is 316 records
Minimum cluster size is  11 records

Sample and Population Counts for Response Variable OKCOHABX
  0:  Sample Count    11503    Population Count 111258567
  1:  Sample Count     1068    Population Count  11449169

R-Square for dependent variable OKCOHABX (Cox & Snell, 1989): 0.003530

-2 * Normalized Log-Likelihood with Intercepts Only :  7796.96
-2 * Normalized Log-Likelihood Full Model           :  7752.50
Approximate Chi-Square (-2 * Log-L Ratio)           :    44.46
Degrees of Freedom                                  :        4

Note: The approximate Chi-Square is not adjusted for clustering.
      Refer to hypothesis test table for adjusted test.
```

## SUDAAN 8.0.2 Output Cont.

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

| Independent Variables and Effects | Beta Coeff. | DEFF Beta #4 | SE Beta | T-Test B=0 | P-value T-Test B=0 |
|---|---|---|---|---|---|
| Intercept | -2.73 | 3.10 | 0.21 | -13.00 | 0.0000 |
| R's age at interview | 0.01 | 3.71 | 0.01 | 1.03 | 0.3064 |
| HIEDUCX | 0.32 | 3.64 | 0.12 | 2.65 | 0.0096 |
| BLACK | 0.27 | 1.53 | 0.11 | 2.55 | 0.0127 |
| FEMALE | 0.06 | 2.81 | 0.10 | 0.61 | 0.5460 |

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

| Contrast | Degrees of Freedom | Wald F | P-value Wald F |
|---|---|---|---|
| OVERALL MODEL | 5 | 542.13 | 0.0000 |
| MODEL MINUS | | | |
| INTERCEPT | 4 | 6.19 | 0.0002 |
| INTERCEPT | 1 | 168.97 | 0.0000 |
| AGER | 1 | 1.06 | 0.3064 |
| HIEDUCX | 1 | 7.03 | 0.0096 |
| BLACK | 1 | 6.48 | 0.0127 |
| FEMALE | 1 | 0.37 | 0.5460 |

Variance Estimation Method: Taylor Series (WR)
SE Method: Robust (Binder, 1983)
Working Correlations: Independent
Link Function: Logit
Response variable OKCOHABX: OKCOHABX

| Independent Variables and Effects | Odds Ratio | Lower 95% Limit OR | Upper 95% Limit OR |
|---|---|---|---|
| Intercept | 0.07 | 0.04 | 0.10 |
| R's age at interview | 1.01 | 0.99 | 1.02 |
| HIEDUCX | 1.38 | 1.08 | 1.75 |
| BLACK | 1.31 | 1.06 | 1.62 |
| FEMALE | 1.06 | 0.87 | 1.31 |

**STATA 8.0 Program**

```
use "mf_livtog.dta"

svyset [pweight=FINALWGT], strata(SEST) psu(SECU)

generate okcohabx=0
replace  okcohabx=1 if OKCOHAB==1

generate hieducx=0 if HIEDUC <=9
replace  hieducx=1 if HIEDUC >9

generate black=0
replace  black=1 if HISPRACE==3

svylogit okcohabx AGER hieducx black FEMALE, deff
```

**STATA 8.0 Output**

```
. svylogit okcohabx ager hieducx black female, deff

Survey logistic regression

pweight:  finalwgt                          Number of obs     =     12571
Strata:   sest                              Number of strata  =        84
PSU:      secu                              Number of PSUs    =       168
                                            Population size   = 1.227e+08
                                            F(   4,     81)   =      5.97
                                            Prob > F          =    0.0003

------------------------------------------------------------------------------
    okcohabx |      Coef.   Std. Err.      Deff
-------------+----------------------------------------------------------------
        ager |   .0072349   .0070302    3.57503
     hieducx |   .3204423   .1208391    3.59665
       black |   .2707356   .1063407   1.532045
      female |   .0626189   .1032816   2.819198
       _cons |  -2.727918   .2098602   3.029195
------------------------------------------------------------------------------
```

# Appendix VIII

## Median Design Effect for Female, Male, and Pregnancy File Means and Proportions

Median design effects for the female respondent data, displayed in table VI, were based on the proportion of women:

1. Whose first menstrual period was before age 13
2. Who had had at least one completed pregnancy
3. Who had at least one live birth
4. Who were fecund
5. Whose current contraceptive method was either the pill or a male condom
6. Who had ever used the pill
7. Who had ever used a male condom
8. Whose first method of contraception was either the pill or a male condom
9. Who intended to have additional children
10. Who strongly agreed with the statement that "A young couple should not live together unless they are married"

Median design effects for the pregnancy data, displayed in table VII, were based on the proportion of babies:

1. Who were not breast feed
2. Who were delivered vaginally
3. Whose delivery was paid through personal income or private insurance

Three additional variables include the proportion of pregnancies in which:

4. The length of the pregnancy exceeded 38 weeks
5. The first baby was male
6. The outcome was a live birth

Median design effects for the male respondent data, displayed in table VIII, were based on the proportion of men:

1. Who were surgically sterile
2. Who had ever had sex
3. Who had ever biologically fathered a child
4. Who had ever cohabited with a woman (excluding wives)
5. Who had ever cohabited (including premarital cohabitation)

The direct estimates of the sampling variances were computed for each of these outcomes using the "with replacement" variance estimator (that is, DESIGN = WR) in SUDAAN. The parameter estimates and the direct variance estimates were computed for the female respondent data file (one record for each of the 7,643 responding women), for the pregnancy database (one or more records for each of the 13,593 responding women with at least one pregnancy), and for the male respondent data file (one record for each of the 4,928 responding men). For each data file, the parameter estimates and their sampling variances were computed for all women, pregnancies, or men and for each of the three race/ethnicity categories (Hispanic, non-Hispanic black, and other women).

**Table VI. Median design effects for 10 female respondent file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6**

| Characteristic | Race and ethnicity | | | |
|---|---|---|---|---|
| | Hispanic | Black | Other | Total |
| Total | 1.52 | 1.55 | 2.31 | 2.31 |
| **Age** | | | | |
| 15–17 years | 1.37 | 1.22 | 1.33 | 1.39 |
| 18–19 years | 1.32 | 1.49 | 1.38 | 1.55 |
| 20–24 years | 1.79 | 1.77 | 1.93 | 2.46 |
| 25–29 years | 1.77 | 1.40 | 1.80 | 1.77 |
| 30–34 years | 1.25 | 1.41 | 1.56 | 1.64 |
| 35–39 years | 1.28 | 1.46 | 1.59 | 1.69 |
| 40–44 years | 1.34 | 1.30 | 1.58 | 1.76 |
| **Marital status** | | | | |
| Married | 1.27 | 1.37 | 1.91 | 1.95 |
| Widowed, divorced, or separated | 1.40 | 1.97 | 1.23 | 1.55 |
| Never married | 1.30 | 1.55 | 1.78 | 2.01 |
| **Education** | | | | |
| Less than high school | 1.59 | 1.48 | 1.38 | 1.71 |
| High school diploma | 1.27 | 1.59 | 1.52 | 1.82 |
| Some college | 1.45 | 2.05 | 2.45 | 2.53 |
| College graduate | 1.63 | 1.74 | 1.95 | 1.95 |
| **Poverty level** | | | | |
| 0–100% | 1.78 | 1.83 | 2.14 | 2.09 |
| 101–200% | 1.10 | 1.40 | 1.88 | 1.98 |
| 201–399% | 1.76 | 1.36 | 1.96 | 2.08 |
| 400% or more | 1.36 | 1.36 | 1.67 | 1.86 |
| **Metropolitan residence** | | | | |
| Metropolitan | 1.47 | 1.38 | 2.23 | 2.24 |
| Nonmetropolitan | 1.59 | 1.43 | 2.26 | 2.33 |
| **Labor force status** | | | | |
| Full-time work | 1.66 | 1.49 | 1.90 | 2.04 |
| Part-time work | 1.45 | 1.53 | 1.75 | 1.97 |
| In school | 1.58 | 1.63 | 1.32 | 1.50 |
| Other | 1.50 | 1.71 | 1.45 | 1.64 |

**Table VII. Median design effects for six female pregnancy file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6**

| Characteristic | Race and ethnicity | | | |
|---|---|---|---|---|
| | Hispanic | Black | Other | Total |
| Total | 2.81 | 2.75 | 3.68 | 3.91 |
| Age at outcome[1] | | | | |
| 15–17 years | 1.77 | 1.84 | 1.47 | 1.84 |
| 18–19 years | 1.41 | 1.43 | 1.79 | 1.98 |
| 20–24 years | 2.02 | 2.24 | 2.62 | 2.87 |
| 25–29 years | 1.94 | 1.93 | 1.97 | 2.29 |
| 30–34 years | 1.74 | 1.97 | 1.87 | 2.06 |
| 35–39 years | 1.55 | 1.67 | 1.91 | 2.28 |
| 40–44 years | 1.20 | 1.05 | 2.18 | 2.41 |
| Marital status at outcome | | | | |
| Married | 2.16 | 2.17 | 2.93 | 3.15 |
| Widowed, divorced, or separated | 2.08 | 2.86 | 2.26 | 2.42 |
| Never married | 3.05 | 2.99 | 2.96 | 3.72 |
| Education[2] | | | | |
| Less than high school | 2.85 | 2.84 | 2.20 | 3.71 |
| High school diploma | 2.39 | 1.96 | 3.52 | 3.75 |
| Some college | 1.70 | 2.40 | 2.43 | 2.35 |
| College graduate | 1.66 | 3.39 | 2.19 | 2.56 |
| Poverty level[2] | | | | |
| 0–100% | 2.82 | 2.55 | 3.92 | 3.55 |
| 101–200% | 1.60 | 3.02 | 2.28 | 2.58 |
| 201–399% | 2.03 | 2.21 | 2.57 | 2.58 |
| 400% or more | 2.06 | 2.75 | 2.08 | 2.44 |
| Metropolitan residence[2] | | | | |
| Metropolitan | 2.88 | 2.71 | 3.33 | 3.89 |
| Nonmetropolitan | 2.20 | 1.27 | 3.40 | 2.89 |
| Labor force status[2] | | | | |
| Full-time work | 2.18 | 2.20 | 3.35 | 3.51 |
| Part-time work | 2.18 | 2.49 | 3.04 | 3.01 |
| In school | 0.39 | 2.04 | 1.23 | 1.63 |
| Other | 2.07 | 3.16 | 2.74 | 3.36 |

[1]Status at the end of the pregnancy interval.
[2]Measured at time of interview.

**Table VIII. Median design effects for five male respondent file variables, by race and ethnicity and demographic characteristics: National Survey of Family Growth, Cycle 6**

| Characteristic | Race and ethnicity | | | |
|---|---|---|---|---|
| | Hispanic | Black | Other | Total |
| Total | 2.1 | 1.23 | 2.02 | 2.31 |
| Age | | | | |
| 15–17 years | 0.87 | 0.83 | 2.29 | 2.16 |
| 18–19 years | 1.36 | 0.94 | 1.57 | 1.52 |
| 20–24 years | 1.84 | 1.45 | 1.30 | 1.77 |
| 25–29 years | 1.70 | 1.41 | 1.91 | 1.87 |
| 30–34 years | 1.33 | 2.15 | 1.26 | 1.31 |
| 35–39 years | 2.37 | 1.83 | 1.83 | 2.32 |
| 40–44 years | 1.79 | 1.14 | 1.65 | 1.81 |
| Marital status | | | | |
| Married | 2.35 | 1.04 | 1.55 | 2.13 |
| Widowed, divorced, or separated | 1.67 | 1.40 | 1.92 | 2.17 |
| Never married | 1.56 | 1.44 | 2.08 | 2.38 |
| Education | | | | |
| Less than high school | 1.46 | 1.16 | 3.00 | 2.66 |
| High school diploma | 2.18 | 1.06 | 1.87 | 2.01 |
| Some college | 1.43 | 1.56 | 2.09 | 2.34 |
| College graduate | 1.38 | 1.45 | 1.79 | 1.86 |
| Poverty level | | | | |
| 0–100% | 1.78 | 1.22 | 2.08 | 1.97 |
| 101–200% | 2.31 | 1.53 | 1.85 | 1.89 |
| 201–399% | 1.66 | 1.70 | 1.88 | 2.11 |
| 400% or more | 1.64 | 1.67 | 2.37 | 2.51 |
| Metropolitan residence | | | | |
| Metropolitan | 2.06 | 1.36 | 2.30 | 2.80 |
| Nonmetropolitan | 1.51 | 0.54 | 1.68 | 1.91 |
| Labor force status | | | | |
| Full-time work | 1.54 | 1.35 | 1.51 | 1.65 |
| Part-time work | 2.61 | 1.76 | 2.63 | 2.25 |
| In school | 0.54 | 0.75 | 2.24 | 2.31 |
| Other | 2.23 | 1.50 | 2.11 | 2.07 |

# Vital and Health Statistics series descriptions

SERIES 1. **Programs and Collection Procedures**—These reports describe the data collection programs of the National Center for Health Statistics. They include descriptions of the methods used to collect and process the data, definitions, and other material necessary for understanding the data.

SERIES 2. **Data Evaluation and Methods Research**—These reports are studies of new statistical methods and include analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. These studies also include experimental tests of new survey methods and comparisons of U.S. methodology with those of other countries.

SERIES 3. **Analytical and Epidemiological Studies**—These reports present analytical or interpretive studies based on vital and health statistics. These reports carry the analyses further than the expository types of reports in the other series.

SERIES 4. **Documents and Committee Reports**—These are final reports of major committees concerned with vital and health statistics and documents such as recommended model vital registration laws and revised birth and death certificates.

SERIES 5. **International Vital and Health Statistics Reports**—These reports are analytical or descriptive reports that compare U.S. vital and health statistics with those of other countries or present other international data of relevance to the health statistics system of the United States.

SERIES 6. **Cognition and Survey Measurement**—These reports are from the National Laboratory for Collaborative Research in Cognition and Survey Measurement. They use methods of cognitive science to design, evaluate, and test survey instruments.

SERIES 10. **Data From the National Health Interview Survey**—These reports contain statistics on illness; unintentional injuries; disability; use of hospital, medical, and other health services; and a wide range of special current health topics covering many aspects of health behaviors, health status, and health care utilization. They are based on data collected in a continuing national household interview survey.

SERIES 11. **Data From the National Health Examination Survey, the National Health and Nutrition Examination Surveys, and the Hispanic Health and Nutrition Examination Survey**—Data from direct examination, testing, and measurement on representative samples of the civilian noninstitutionalized population provide the basis for (1) medically defined total prevalence of specific diseases or conditions in the United States and the distributions of the population with respect to physical, physiological, and psychological characteristics, and (2) analyses of trends and relationships among various measurements and between survey periods.

SERIES 12. **Data From the Institutionalized Population Surveys**—Discontinued in 1975. Reports from these surveys are included in Series 13.

SERIES 13. **Data From the National Health Care Survey**—These reports contain statistics on health resources and the public's use of health care resources including ambulatory, hospital, and long-term care services based on data collected directly from health care providers and provider records.

SERIES 14. **Data on Health Resources: Manpower and Facilities**—Discontinued in 1990. Reports on the numbers, geographic distribution, and characteristics of health resources are now included in Series 13.

SERIES 15. **Data From Special Surveys**—These reports contain statistics on health and health-related topics collected in special surveys that are not part of the continuing data systems of the National Center for Health Statistics.

SERIES 16. **Compilations of Advance Data From Vital and Health Statistics**—Advance Data Reports provide early release of information from the National Center for Health Statistics' health and demographic surveys. They are compiled in the order in which they are published. Some of these releases may be followed by detailed reports in Series 10–13.

SERIES 20. **Data on Mortality**—These reports contain statistics on mortality that are not included in regular, annual, or monthly reports. Special analyses by cause of death, age, other demographic variables, and geographic and trend analyses are included.

SERIES 21. **Data on Natality, Marriage, and Divorce**—These reports contain statistics on natality, marriage, and divorce that are not included in regular, annual, or monthly reports. Special analyses by health and demographic variables and geographic and trend analyses are included.

SERIES 22. **Data From the National Mortality and Natality Surveys**—Discontinued in 1975. Reports from these sample surveys, based on vital records, are now published in Series 20 or 21.

SERIES 23. **Data From the National Survey of Family Growth**—These reports contain statistics on factors that affect birth rates, including contraception, infertility, cohabitation, marriage, divorce, and remarriage; adoption; use of medical care for family planning and infertility; and related maternal and infant health topics. These statistics are based on national surveys of women of childbearing age.

SERIES 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**—These include advance reports of births, deaths, marriages, and divorces based on final data from the National Vital Statistics System that were published as *National Vital Statistics Reports* (NVSR), formerly *Monthly Vital Statistics Report*. These reports provide highlights and summaries of detailed data subsequently published in *Vital Statistics of the United States*. Other special reports published here provide selected findings based on final data from the National Vital Statistics System and may be followed by detailed reports in Series 20 or 21.

For answers to questions about this report or for a list of reports published in these series, contact:

Information Dissemination Staff
National Center for Health Statistics
Centers for Disease Control and Prevention
3311 Toledo Road, Room 5412
Hyattsville, MD 20782

1-866-441-NCHS (6247)
E-mail: nchsquery@cdc.gov
Internet: www.cdc.gov/nchs

**U.S. DEPARTMENT OF**
**HEALTH & HUMAN SERVICES**

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road
Hyattsville, MD 20782

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE, $300

MEDIA MAIL
POSTAGE & FEES PAID
CDC/NCHS
PERMIT NO. G-284

05-0387 (7/2006)
CS103369
T25968
DHHS Publication No. (PHS) 2006-1342, Series 2, No. 142