SAFER • HEALTHIER • PEOPLE™

Monitoring the Nation's Health

Vital and Health Statistics

Series 2, Number 158

June 2013

Responsive Design, Weighting, and Variance Estimation in the 2006–2010 National Survey of Family Growth

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

Suggested citation

Lepkowski JM, Mosher WD, Groves RM, et al. Responsive design, weighting, and variance estimation in the 2006–2010 National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 2(158). 2013.

Library of Congress Cataloging-in-Publication Data

Responsive design, weighting, and variance estimation in the 2006–2010 National Survey of Family Growth.
p. ; cm.— (Vital and health statistics. Series 2, Data evaluation and methods research ; no. 158) (DHHS publication ; no. (PHS) 2013-1358)
"June 2013."
Includes bibliographical references and index.
ISBN 0-8406-0658-3 (alk. paper)
I. National Center for Health Statistics (U.S.) II. Series; Vital and health statistics. Series 2, Data evaluation and methods research ; no. 158. III. Series: DHHS publication ; no. (PHS) 2013-1358. 0276-4733
[DNLM: 1. National Survey of Family Growth (U.S.) 2. Data Collection—United States. 3. Family Characteristics—United States. 4. Analysis of Variance—United States. 5. Data Interpretation, Statistical—United States. 6. Research Design—United States. W2 A N148vb no.158 2013]
RA407.3
614.4'273—dc23
2013013428

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 158

Responsive Design, Weighting, and Variance Estimation in the 2006–2010 National Survey of Family Growth

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
June 2013
DHHS Publication No. 2013–1358

Contents

Text Figures

Text Tables

Appendix Figures

Appendix Tables

Acknowledgments

The 2006–2010 National Survey of Family Growth (NSFG) was designed and conducted by the Centers for Disease Control and Prevention's (CDC) National Center for Health Statistics (NCHS) and its 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 of ISR, in consultation with William Mosher and Karen Davis of NCHS. Appendix III of this report was written by Brady T. West of ISR. The information in this report is based on survey design documentation prepared by ISR staff and on internal NCHS memoranda.

The 2006–2010 NSFG was jointly planned and funded by the following programs and agencies of the U.S. Department of Health and Human Services:

- Eunice Kennedy Shriver National Institute for Child Health and Human Development
- Office of Population Affairs
- National Center for Health Statistics, CDC
- Division of HIV/AIDS Prevention, CDC
- Division of Sexually Transmitted Disease Prevention, CDC
- Division of Reproductive Health, CDC
- Children's Bureau of the Administration for Children and Families (ACF)
- Office of Planning, Research, and Evaluation, ACF
- Office of the Assistant Secretary for Planning and Evaluation

NCHS gratefully acknowledges the contributions of these programs and agencies, and all others who assisted in designing and implementing the NSFG.

The authors of this report gratefully acknowledge the comments of Van L. Parsons, Ph.D., of NCHS for his peer review of the manuscript, and the comments of Anjani Chandra, Ph.D., and Casey Copen, Ph.D., on the section on "Using Survey Estimation Software." This report was prepared under the general direction of Charles J. Rothwell, Acting Director of the National Center for Health Statistics, and Stephanie J. Ventura, Chief of the Reproductive Statistics Branch of NCHS' Division of Vital Statistics. The report was produced by CDC/OSELS/NCHS/OD/Office of Information Services, Information Design and Publishing Staff: Danielle Woods edited the report, typesetting was done by Jacqueline M. Davis, and graphics were produced by Sarah Hinkle.

Abstract

Objective

The National Survey of Family Growth (NSFG) collects data on pregnancy, childbearing, men's and women's health, and parenting from a national sample of men and women aged 15–44 in the United States. The 2006–2010 NSFG design was a significant departure from the previous periodic design, used in 1973–2002. This report shows fieldwork results and weighting, imputation, and variance estimation procedures. The report should be useful to users of the 2006–2010 public-use data file and to survey methodologists wishing to learn how the NSFG was conducted.

Methods

NSFG's new design is based on an independent national probability sample of men and women aged 15–44. The University of Michigan's Institute for Social Research conducted fieldwork under a contract with the National Center for Health Statistics. Professional female interviewers conducted in-person, face-to-face interviews using laptop computers. A responsive design approach was used in planning and managing the fieldwork for NSFG to control costs and reduce nonresponse bias.

Results

The 2006–2010 NSFG is based on 22,682 completed interviews—10,403 interviews with men and 12,279 with women. Interviews with men lasted an average of 52 minutes, and for women, 71 minutes. Weighted response rates were 75% among men, 78% among women, and 77% overall.

Analysis of NSFG data requires the use of sampling weights and estimation of sampling errors that account for the complex sample design and estimation features of the survey. Sampling weights are provided on the data files. The rate of missing data in the survey is generally low.

Keywords: survey methodology • response rates • paradata • survey management

Responsive Design, Weighting, and Variance Estimation in the 2006–2010 National Survey of Family Growth

by James M. Lepkowski, Ph.D., Institute for Social Research, University of Michigan; William D. Mosher, Ph.D., National Center for Health Statistics; Robert M. Groves, Ph.D., Brady T. West, Ph.D., James Wagner, Ph.D., and Haley Gu, Ph.D., Institute for Social Research, University of Michigan

Executive Summary

The National Survey of Family Growth (NSFG) is designed to provide national statistics on factors affecting childbearing, marriage, and parenting from a national probability sample of women and men aged 15–44. The 2006–2010 NSFG used a responsive design approach to manage the field work of the survey, to control costs, to oversample teenage and minority groups, and to reduce bias in the resulting sample. This approach represented a significant change in the design, methodology, and procedures of the survey. These changes in methods and the extended 4-year interviewing period required the release of two previous technical reports that described how the survey was planned and designed (1,2). This report describes the *outcomes* of the fieldwork and responsive design and the corresponding procedures used for weighting, imputation, and variance estimation in the 2006–2010 survey. This information should be useful for those who intend to do statistical research with NSFG data, and for survey methodologists who want to compare their procedures with those used in the NSFG.

NSFG is designed and administered by the National Center for Health Statistics (NCHS), an agency of the Centers for Disease Control and Prevention, 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, and infertility.
- Medical care associated with contraception, infertility, and childbirth.
- Factors affecting marriage, divorce, cohabitation, and adoption.
- What fathers do to help raise their children.
- Men's and women's attitudes about childbearing, parenthood, and marriage.

The 2006–2010 NSFG was conducted by the University of Michigan's Institute for Social Research (ISR) under a contract with NCHS. The 2006–2010 NSFG was the first time the NSFG was fielded using a *continuous design*, meaning that NSFG interviewing was conducted 48 weeks per year over a 4-year period, instead of completing interviewing in 8–12 months.

Interviewing for the 2006–2010 survey began in late June 2006 and ended in June 2010—a 48-month period. Interviews were conducted with a national probability sample of women and men aged 15–44 living in households in the United States. The interviews were administered *in person* by trained female interviewers using a laptop or notebook computer, a procedure called computer-assisted personal interviewing (CAPI; see Appendix I). The interviews for women lasted an average of 71 minutes and the interviews for men lasted an average of 52 minutes. Interviews were conducted in English and Spanish. About one-third of interviews with Hispanic persons, or about 7% of all interviews, were conducted in Spanish by bilingual interviewers.

The 2006–2010 NSFG was based on a sampling plan that was intended to provide larger samples at a lower cost per interview. Reducing costs and increasing the predictability of costs were important goals of this design, and both goals were accomplished. It was also important to obtain large samples of black, Hispanic, and teen (aged 15–19) respondents, and this goal was achieved also.

The 2006–2010 national sample was drawn from 110 major areas, or primary sampling units (PSUs; see Appendix I), divided into four national subsamples. Interviewing was done for 1 year in each of the four subsamples, so the entire 110-PSU design could be completed in a 4-year period. The entire 4-year data file has 22,682 interviews in 110 PSUs—the largest sample in NSFG's history. The response rate (see Appendix I) was 77% overall—78% for women, 75% for men, and 77% for male and female teenagers.

Interviewing for NSFG was divided into four 12-week quarters each year. Each 12-week quarter was divided into two *phases*: the first 10 weeks were called "phase 1." During this time, a $40 incentive, or token of appreciation, was used. During weeks 11 and 12, "phase 2" procedures were implemented, in which a random sample of one-third of the remaining nonresponding cases was retained in the sample. This reduced sample received 3 times as much interviewer effort per case, and the incentive was doubled to $80; these changes were intended to raise response rates and correct any bias due to nonresponse. This kind of sample is sometimes called a "two-phase" or "double sample" design. (See Appendix I for definitions.)

Sampling weights were used to compensate for the different sampling rates of these various groups and for different nonresponse rates. This report shows how sampling errors should be estimated—using software that takes into account the weights and the stratified cluster sample design (see "Strata and stratification" definition in Appendix I). Software packages such as SAS, Stata, SPSS, and SUDAAN have procedures that will calculate sampling errors in this way.

For about 650 key variables, referred to in this and other NSFG reports as "recodes" (see Appendix I), item missing values have been replaced in the data file by predicted or imputed values. The imputed values are identified through a companion variable or "imputation flag" to indicate whether the value for a particular case was imputed or reported. The imputation rates for most of these recodes are very low—usually less than 1.0%. The highest imputation rate was for income, with 10.8% imputed.

The text of this report describes responsive design aspects and outcomes of the fieldwork, as well as weighting, imputation, and variance estimation procedures used with the 2006–2010 NSFG. This report also includes three appendices:

- Appendix I defines key technical terms used in the report.
- Appendix II describes how NSFG has experimented with incentives to improve survey data collection.
- Appendix III answers a common user question about whether variance estimation needs to account for more than the first-stage sampling units when sampling variances are estimated.

Although *plans* for some of the innovations in NSFG were described in two previous reports (1,2), this report describes the *results* of the new features of the 2006–2010 design. These features may be of particular interest to survey methodologists, including:

- A sample design with a small interviewing staff and rotating sampled areas (PSUs).
- Interviewer recruitment and assignment to a fixed 30-hour work week.
- Continuous, real-time survey management with paradata to ration and allocate interviewer effort and control costs.
- Randomized experimental interventions to improve response rates in phase 1 of fieldwork (weeks 1–10 of each 12-week quarter).
- The use of paradata to select and manage a two-phase sample in weeks 11–12 of each 12-week quarter.
- Experimental evaluation of the use of incentives during the second phase of fieldwork.
- Nonresponse adjustment and weighting using paradata and conventional demographic variables.
- Imputation of missing data using sequential regression procedures.
- Thorough trimming procedures to control pronounced variability in sampling weights.

2006–2010 NSFG Objectives

NSFG was established at NCHS in 1971 and Cycle 1 was conducted in 1973 to continue surveys previously conducted by other organizations (3–7). NCHS has conducted NSFG seven times since 1973. In each of the first six NSFG surveys, survey interviewing was completed in 1 year or less. For a summary of NSFG's history, see Groves et al. (1).

The results of the 2002 NSFG were used to inform the design of the continuous NSFG, in which NSFG interviewing would be done every year indefinitely, as long as funding and other circumstances permit. The fieldwork plan for the 2006–2010 sample was to complete interviewing in 4 years. After that, another sample

would be drawn for the next few years of interviewing.

The principal goal of the 2006–2010 design was to increase sample sizes substantially, especially for teenagers, black men and women, and Hispanic men and women, and to do so at a lower cost per case.

Several factors made this goal seem difficult, including uncertainties about eligibility rates in samples of U.S. addresses; uncertainties about contact rates and cooperation rates; and costs of recruiting, training, and managing a large staff of short-term, one-time employees.

For example, the 2002 NSFG used 246 interviewers to conduct 12,571 interviews over a 12-month period (or only 51 interviews per interviewer). The 2002 NSFG design posed several challenges. The interviewers did not have adequate time to practice tasks before the survey ended and large investments in training were wasted once the survey ended. This process decreased efficiency and increased costs per interview, meaning that quality control with this large number of interviewers working for a short time was more difficult than if the interviewers had a permanent assignment and closer supervision.

Further, interviewers have traditionally been employed part time, and toward the end of a study did not have enough hours to work. As a result, they tended to leave the project before data collection ended to find jobs with more hours and thus, more pay. A steady flow of work and a more even distribution of hours was expected to lead to greater efficiency and fewer costly staff losses at a critical time in data collection.

To produce more predictable results and to better control costs, the 2006–2010 NSFG conducted 22,682 interviews with 113 interviewers, or an average of 201 interviews per interviewer, compared with 51 per interviewer in 2002—4 times as many per interviewer. This was more cost efficient because interviewers learned and perfected skills that could be performed more efficiently over an extended period of time. The field organization aspects of the design are discussed further in the ''Organization of Fieldwork'' section.

NSFG provides data needed by several federal programs in addition to NCHS (1; see also ''Acknowledgments''). During consultations, these agencies expressed the need to collect more interviews more often and to release the results more frequently while controlling costs (1). The design described here accomplished these goals.

Detailed design specifications for the survey were published previously (2). In brief, the NSFG's 2006–2010 sample design was based on the following objectives:

- The target population (see Appendix I) for the survey was the population of men and women aged 15–44 in households in the 50 U.S. states and the District of Columbia (DC). Finding this population required short screening interviews to determine if anyone aged 15–44 lived in the household, and if so, required selecting one person for the interview.
- Interviewing had to be conducted in person by well-trained female interviewers because of the complexity and sensitivity of the interview. Signed parental consent was required for those aged 15–17.
- Questionnaires and interviews were to be available in English and Spanish. Interviews were expected to last an average of 80 minutes for women and 60 minutes for men.
- Available funds supported a sample of about 5,000 interviews per year for 4 years, yielding about 20,000 interviews over a 4-year data collection period. The sample should consist of about 45% male respondents and 55% female respondents, and about 20% Hispanic respondents, 20% black respondents, and 20% teen respondents.

Given the limited funding available to meet the objectives of NSFG, a new design was needed that provided greater cost efficiency and greater control over the costs and results of the survey fieldwork.

NSFG staff and the NSFG contractor, the University of Michigan's ISR, proposed a design with the following features:

- A stratified, multistage sample design allocated to give all interviewers a workload large enough to maximize their efficiency.
- Interviews would be conducted 48 weeks per year for several years, thus writing off the costs of hiring and training interviewers over a longer period of time, retaining a higher proportion of interviewing staff, and maintaining the benefits of an increasingly experienced field staff.
- A relatively small number of PSUs at any one time to permit a smaller, more closely supervised interviewing staff, with PSUs that would be rotated annually to assemble a more precise sample for national estimation over time. Previous cycles of the NSFG using large numbers of PSUs had been costly despite other cost-saving features (8). (A few large PSUs would stay in the sample every year, but most would rotate annually; therefore most interviewers would work for 1 full year, but some would work up to 4 years on the project.)
- Increased use of paradata (data about the data collection process) to allow survey managers to adjust interviewer effort to achieve higher response rates and to keep the data collection cost efficient. This approach to data collection is called ''responsive'' or ''adaptive'' survey design (9).

Summary of Sample Design

Overall, the NSFG sample design consisted of *five stages* of selection: PSUs, blocks or segments, housing units, one eligible person per housing unit, and housing units or persons for phase 2 data collection (see ''Double sample'' in Appendix I).

Primary Sampling Unit Selection

The sample of 110 PSUs was divided into four fully representative national samples for the 2006–2010 NSFG. The four annual national samples allow new samples to be introduced in each of 4 consecutive years of data collection. These four national samples allowed ISR to make changes, as required, to data collection, survey questions, and other design features once a year. The samples also permitted the survey to be conducted with a smaller, better-managed staff and still have nationally representative samples that could be combined into a large 4-year sample. Lepkowski et al. discuss the details of how PSUs were selected (2).

Block or Segment Selection

In the second stage of selection, U.S. Census Bureau land areas known as census blocks were used to further divide the land area of each of the 110 sample PSUs. Census blocks were grouped into domains (see Appendix I), and within each domain they were chosen with probabilities proportionate to an estimated number of 2000 Census occupied housing units (2).

The second-stage sampling units were single census blocks, or combinations of census blocks, that had sufficient numbers of households to sustain efficient survey data collection. For efficient data collection, geographically contiguous blocks were linked to one another into units called segments, with at least 75 households in urban areas or 50 households in rural areas. These segments made up the second-stage sampling units.

Before linking, blocks were grouped into one of four domains within each PSU based on the percentage of black or Hispanic households in the segment:

1. Non-minority
2. More than 10% of black persons but less than 10% of Hispanic persons
3. More than 10% of Hispanic persons but less than 10% of black persons
4. More than 10% of black and more than 10% of Hispanic persons.

Sampling rates for households within each of these four domains were determined in order to achieve target numbers of completed interviews with black and Hispanic households. Higher rates were needed for domains 2–4 in order to achieve target subgroup sizes. After simulation of design alternatives and examination of sample sizes in target groups and effective sample sizes, sampling rates for domains 2–4 were set 200% to 250% higher than the rate for domain 1.

In the 2006–2010 NSFG, exactly 12 segments were selected in each of 80 of the nonmetropolitan, nonself-representing (NSR) sample PSUs. The 28 largest PSUs received an allocation of segments that was proportionate to size, with the smallest PSUs receiving approximately 12 segments and the largest more than twice as many.

Segments were selected within a PSU using a systematic selection 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.

Within a PSU, one-quarter of the segments allocated to each PSU in the yearly sample were randomly selected in each 12-week data collection quarter. If there were exactly 12 segments in each of the 33 PSUs in a yearly sample, then three would be used in each quarter, so there would be an expected 3 x 33 = 99 segments in a calendar quarter sample. However, because the eight largest and the five metropolitan NSR PSUs had on average more than 12 segments due to their larger populations, the number of segments in a calendar quarter was approximately 116 segments. Over the entire year, approximately 464 segments were in the 2006–2010 NSFG annual national sample.

Table A presents counts of the number of segments selected and listed, and the varying sampling rates applied to each domain. An average of 116 segments were selected and sent for listing in each of the 16 quarters. The sampling rates across domains varied from 1 in 267 in domain 4 (neighborhoods with high proportions of black and Hispanic households) to 1 in 956 in domain 1 (neighborhoods with few black and Hispanic households)—a factor of about 3.5 to 1.0. The net result of this variation was an increase in the number of black and Hispanic respondents in the sample, relative to the proportionate distribution seen in the population.

Table B shows that 1,840 segments were selected across the 4 years of data collection in the 110 PSUs. Thus, on average, about 115 segments were listed each quarter, or about 460 per year.

Housing Unit Selection

Lepkowski et al. (2) offer a detailed account of housing unit selection, whereas this section describes some of the results of the application of those procedures in the 2006–2010 NSFG. Housing unit lists were prepared for each selected segment. These segment lists came from one of three sources:

- Unused housing unit listings from the 2002 NSFG (Cycle 6)
- Addresses for housing units obtained from a commercial vendor of the U.S. Postal Service's Delivery Sequence File (DSF; see Appendix I)
- A field listing of addresses created by interviewers visiting segments before data collection began (also known as a "scratch" listing)

A primary parameter in the allocation of housing units to the PSU and segment levels was interviewer workload. Interviewers were recruited and hired to work an average of 30 hours per week, or approximately 360 hours in a 12-week quarter. This allocation led to variation in probabilities of selection of housing units across segments within and among PSUs. The variation was compensated for in the weighting process (described below), although the added variability in sample weights from varying line probabilities at the segment level had the potential to increase the variability of survey estimates (see "Weighting Procedures" section of this report).

Table A. Average number of segments allocated, average number of segments listed, and average sampling rate, by domain: National Survey of Family Growth, 2006–2010

Domain	Average number of segments allocated (range)	Average number of segments listed (range)	Average within-domain sampling rate (range)
Total	115.9 (110–122)	115.1 (108–121)	1 in 488 (1 in 188 to 1,210)
1. Non-minority persons	38.4 (30–47)	38.4 (30–47)	1 in 956 (1 in 749 to 1,210)
2. More than 10% black persons	31.1 (21–39)	30.6 (21–39)	1 in 325 (1 in 188 to 496)
3. More than 10% Hispanic persons	26.8 (21–36)	26.6 (21–36)	1 in 404 (1 in 246 to 517)
4. More than 10% black persons and more than 10% Hispanic persons	19.6 (18–23)	19.4 (17–23)	1 in 267 (1 in 224 to 352)

Table B. Segments and housing units selected for phase 1 and phase 2 samples in all 16 quarters of the National Survey of Family Growth, 2006–2010

Phase	Number	Average per quarter
Phase 1 (weeks 1–10)		
Segments listed	1,840	115
Active main addresses, week 10	9,066	567
Active screener addresses, week 10	8,082	505
Phase 2 (weeks 11–12)		
Segments selected	1,224	77
Main interview addresses selected	2,896	181
Screener addresses selected	2,368	148

Table C presents the number of addresses selected and the number of those addresses that were screened and found to have one or more eligible persons. Over 78,000 addresses were selected across the 16 quarters of data collection, or 4,880 on average per quarter. Of these, an average of 2,008 per quarter contained an eligible person aged 15–44. One eligible person was selected from these addresses, and an average of 1,418 interviews per quarter was completed.

As shown in Table A, the average number of segments selected and listed nationwide was 115 in a 12-week quarter. Interviewers were equipped with a sample management system (SurveyTrak) on a laptop computer that contained all addresses in each segment. Non-minority segments were selected at the lowest rate (1 in 956 on average, see Table A), compared with a rate of 1 in 267 for segments with more than 10% black and more than 10% Hispanic populations.

Person Selection

Once housing units were selected, interviewers attempted a screener interview for each occupied housing unit in the sample. Screening consisted of a short questionnaire administered at the doorstep to determine whether any persons aged 15–44 resided in the occupied housing unit (see definition of "Eligible household" in Appendix I).

Within selected housing units with one or more persons aged 15–44, one eligible person was selected at random to be interviewed. When interviewers visited selected housing units in assigned segments, they completed a household roster of all persons who usually lived there. The details of the within-household selection procedure are discussed elsewhere (2), and Figure 1 illustrates the process. Size values were assigned to sample persons by age, sex, and race and ethnicity, as shown in the upper panel of the figure. Larger sizes were assigned to teenagers and women to increase the number of each group in the final sample. In the lower panel, each person in a hypothetical household was assigned a measure of size value corresponding to the upper panel. (Persons outside the 15–44 age range were assigned a size value of zero.) The

Table C. Number of selected addresses, screened eligible households, and main interviews, by data collection release, and average number of addresses, eligible households, and main interviews per quarter: National Survey of Family Growth, 2006–2010

Characteristic	Number
Selected addresses[1]	
Total	78,082
Average per quarter	4,880
Screened eligible households[2]	
Total	32,134
Average per quarter	2,008
Main interviews[3]	
Total	22,682
Average per quarter	1,418

[1]Selected addresses are the number of addresses selected into the screener sample.
[2]Screened eligible households are screener addresses with final screener result of completed eligible interview.
[3]Main interviews are screened eligible households with a completed interview with the selected respondent.

Age and sex	Black	Hispanic	White or other
Female			
15–19	1.00	1.00	0.84
20–44	0.57	0.56	0.26
Male			
15–19	0.98	0.98	0.82
20–44	0.53	0.52	0.22

Illustration:

Race and ethnicity	Sex	Age	Measure of size	Cumulative measure of size	Random number (from 0 to 2.63)
Black	Female	6	0.00	–	
Black	Female	15	1.00	1.00	← 0.95
Black	Female	40	0.57	1.57	
Black	Male	10	0.00	–	
Black	Male	21	0.53	2.10	
Black	Male	42	0.53	2.63	

– Quantity zero.

Figure 1. Within-household measures of size and an illustration of within-household selection in the 2006–2010 National Survey of Family Growth

size values were then cumulated, and a random number from zero to the sum of size values was generated (using the Blaise system in the CAPI application). In the illustration, the random number is 0.95. This number was compared with the cumulated sizes, and the person with cumulated size which first exceeded the random number was selected. As a result, in Figure 1, the black woman aged 15 with an accumulated size value of 1.00 had a cumulated size that first exceeded the random number, and she was chosen.

Responsive Design and Management of Fieldwork

The majority of survey expense in an in-person, area-probability survey like NSFG is in data collection, particularly the training of interviewers and their labor and travel during data collection. Improving the cost-efficiency of an in-person survey must focus on interviewer training and labor. To control survey costs, improvements in basic productivity measures such as output (interviews) per unit of labor (interviewer hours) must be sought in times of reduced funding for survey data collection.

In considering how to improve the cost efficiency of the 2006–2010 data collection, NSFG staff at NCHS and ISR determined that data indicators of interviewer performance in forms that could be readily analyzed were limited. Data that could inform survey managers about interviewer effort, the daily status of each sample case, or other measures of survey progress were not being used. These kinds of data are generated as the interviewer does her work, and are recorded in parallel with survey interviews. But systems to access and use these paradata had not been available for use in managing surveys *during data collection* until recently. Yet, staff recognized that improvements in data collection efficiency could be made if such data could be examined and used to optimize interviewer effort during ongoing data collection.

Staff collected evidence and anecdotal reports that strongly suggested cost-efficiency improvements were possible if interviewer behavior could be altered. Staff needed data that could be used to build indicators of system performance systematically throughout the data collection period and needed a willingness to alter small and large data collection features during data collection to try to improve efficiency.

This strategy was risky, considering that procedural changes might not lead to improvement in system efficiency, but could reduce it. And implementing the strategy on an ad hoc basis, without careful thought given to analyzing change in indicators before and after an intervention, could lead to ineffective changes and poorer efficiency. NCHS and ISR staff concluded that procedural changes should, when possible, be made in conjunction with experimental designs to assess the effectiveness of changes in outcomes and indicators.

The strategy of using paradata to identify interventions and then to evaluate the effects of changes requires identification of what can be adjusted to improve efficiency. As part of the planning of the 2006–2010 data collection, NCHS and ISR staff discussed what kinds of interviewer behavior might be monitored to understand where the data collection process could be altered to lead to improvements.

Staff considered many interviewer behaviors that are poorly understood but that also are likely to decrease survey efficiency, including:

- How interviewers allocate their time between conducting screeners, interviews, and other activities.
- How interviewers could make administrative tasks smaller parts of their jobs.
- Whether interviewers working longer rather than shorter work weeks might lead to better allocation of time across tasks.
- Whether interviewers working across several studies at the same time, as in many survey organizations, might be less efficient than interviewers working more hours, weeks, or months on one survey.
- Whether distributing the sample differently across segments might reduce travel time per completed interview.
- Whether study managers could provide interviewers with information that would help them

obtain interviews as efficiently as possible.

The 2006–2010 NSFG studied these behaviors using paradata and formulated interventions that could alter interviewer behavior and improve survey efficiency. The approach was a type of ''responsive design,'' as described by Groves and Heeringa (9) and by Groves et al. (1).

Responsive Survey Design

The responsive survey design considered the following procedures:

- Identifying, before fieldwork begins, a set of design features potentially affecting costs and errors of survey estimates.
- Monitoring those indicators on a daily basis.
- Altering survey procedures during data collection to increase quality and reduce costs and errors.
- Combining survey data, when necessary, from before and after interventions or phases of data collection into a single estimate.

For responsive design to be most effective, survey procedures must be altered during data collection. But altering survey procedures during collection should not be done without prior planning and agreement that it is necessary and likely to lead to improvements in outcomes. Anticipating such changes facilitates intervention and allows study staff to respond or adapt to changing field conditions. This responsive design framework in NSFG was, primarily, a means to decide how to make those alterations and to how to evaluate them.

Responsive design features are attractive for a survey like the 2006–2010 NSFG because they may help manage the uncertainty of some of the key determinants of the final product—the number of completed interviews. Over the course of 4 years of data collection, response rates and the factors that affect them may change. Leaving the design unaltered in the face of these changes could lead to growing inefficiency over time.

Three factors affecting the design and response rates were particularly important in the cost-efficiency of the 2006–2010 NSFG: occupancy rates, eligibility rates, and nonresponse levels. Initial values of these rates and levels were needed to establish basic requirements, such as how many addresses to select per segment for a given interviewer. Initial values would be set on the basis of past data, and as the sample changed from quarter to quarter or from year to year, the initial predictions about occupancy, eligibility, and nonresponse could be inaccurate. The survey needed to generate indicators to monitor these variables, and to be ready to alter the design on the basis of updated values.

In order to develop indicators needed for effective monitoring and intervention, interviewer observational data were generated and administrative systems data were extracted to build indicators (1,2,9,10). The paradata in the 2006–2010 NSFG included:

- Whether a structure appeared to be abandoned or unoccupied.
- The extent of commercial, church, school, and other nonresidential use in the neighborhood.
- Access impediments to a unit (e.g., locked entrance or doorkeeper).
- Observed presence of children under age 15 years in the household.
- Time of day, day of week, and outcome for each visit.
- Hours charged and type of tasks performed during those hours.
- Mileage charges for travel to and from sample segments.
- Whether the householder asked a question (e.g., ''How did you choose my house?'') during a contact.
- Whether the interviewer believes the respondent selected from a screening interview is married to or cohabiting with an opposite-sex partner.

The ''Organization of Fieldwork'' section below describes general interviewing organization and procedures and the results of the process for the 2006–2010 survey. Although many of the results presented are reports of routine features of the survey such as interview length and number of interviewers, the results also contain data related to the responsive design. The 2006–2010 NSFG also had two responsive design features as integral components of the design that utilized these data: interventions to improve response rates and two-phase sampling for nonresponse. The outcomes of both of these features are discussed in the following sections.

Organization of Fieldwork

This section provides data about the basic organization of the fieldwork, how that organization facilitated responsive design, and how the responsive design features led to changes in survey efficiency during the 4 years of data collection.

The 2006–2010 NSFG involved two complex computer-assisted personal interview (CAPI) questionnaires: one for men and one for women. These instruments were written in the Blaise software system version 4.4 (http://www.westat.com/blaise; see Appendix I). Both of these instruments were used in the 2002 NSFG, with some modifications for the 2006–2010 survey.

Table D shows the length of interview for these questionnaires. The female questionnaire required 71 minutes on average, and the male questionnaire required 52 minutes. Table D also shows that both questionnaire administration times increased consistently in length as the age of the respondent increased, as older respondents typically have more births, marriages, cohabitations, and other events to report.

Interviewer productivity is reduced and survey costs are increased by the inevitable learning curves that new interviewers experience in the first months of work. In the 2002 NSFG, 246 interviewers were trained and conducted 12,571 interviews—about 51 per interviewer—in the 12 months of fieldwork (March 2002–February 2003). In contrast, the 2006–2010 NSFG used about 40 interviewers in any given year because the sample of 110 PSUs was divided into four annual replicates (see Appendix I for definition) containing about 35 PSUs each. This smaller group of interviewers worked consistently over the entire year, with workloads designed to maximize their productivity. This design allowed the average interviewer

Table D. Mean length of interview for completed female and male interviews, by age group: National Survey of Family Growth, 2006–2010

Sex and age group	Mean length of interview in minutes
Total	62.3
Female	
Total	71.2
15–19 years	52.7
20–24 years	71.3
25–44 years	76.6
Male	
Total	51.8
15–19 years	42.1
20–24 years	50.5
25–44 years	55.9

NOTE: Calculation of mean length excludes interviews from which timings could not be calculated because of missing data.

Table E. Interviewers trained, by data collection year: National Survey of Family Growth, 2006–2010

Data collection year	Number of interviewers
Year 1 (June 2006–June 2007)	46
Year 2 (June 2007–June 2008)	22
Year 3 (June 2008–June 2009)	26
Year 4 (June 2009–June 2010)	19
Total	113

Table F. Number of completed interviews, by data collection year and race and ethnicity: National Survey of Family Growth, 2006–2010

Race and ethnicity	All years	Data collection year 1 (2006–2007)	2 (2007–2008)	3 (2008–2009)	4 (2009–2010)
Total	22,682	5,555	5,161	5,783	6,183
Black	4,411	969	1,040	1,088	1,314
Hispanic	4,889	1,161	963	1,456	1,309
Other[1]	12,382	3,425	3,158	3,239	3,560

[1]Includes white, Asian, American Indian, and other races.

Table G. Number of completed interviews, by race and ethnicity, sex, and age group: National Survey of Family Growth, 2006–2010

Sex and age group	Race and ethnicity Total	Black	Hispanic	Other
Total	22,682	4,411	4,889	13,382
Male	10,403	1,854	2,297	6,252
Female	12,279	2,557	2,592	7,130
15–19 years	4,662	932	1,035	2,695
Male	2,378	470	551	1,357
Female	2,284	462	484	1,338
20–44 years	18,020	3,479	3,854	10,687
Male	8,025	1,384	1,746	4,895
Female	9,995	2,095	2,108	5,792

to produce considerably more interviews: in 2006–2010, 113 interviewers produced 22,682 interviews (Tables E and F), or 201 per interviewer. This means that the cost of recruiting and training each interviewer was spread out over many more interviews in 2006–2010 than in 2002.

Also note in Table E that the number of interviewers trained in years 2, 3, and 4 was about one-half as large (about 20) as in year 1 (46). This was the case because some first-year interviewers in larger PSUs continued to work in those PSUs since those large metropolitan areas remained in the sample; a few other interviewers worked in a nearby PSU or served as traveling interviewers.

Table F presents the number of completed interviews by data collection year and race and ethnicity, whereas Table G shows the number of completed interviews by sex, age group (ages 15–19 compared with ages 20–44), and race and ethnicity (Hispanic, black, and other; see Appendix I for definition of "race"). As shown in Table F, the target of 5,000 interviews was exceeded in each data collection year, with annual sample sizes ranging from 5,161 in year 2 to 6,183 in year 4.

In 2006–2010, 22,682 interviews were completed—10,403 with men and 12,279 with women (Table G). Interviews with women outnumbered those with men because the sample was designed to be 55% women and 45% men. The 2006–2010 NSFG had 4,411 black and 4,889 Hispanic respondents, more than had ever been interviewed in prior NSFGs. The sample of teenagers aged 15–19 was also the largest sample of teenagers ever included in NSFG—4,662 interviews, including 932 with black teenagers and 1,035 with Hispanic teenagers. This large sample is important because black and Hispanic teenagers have markedly different birth and pregnancy rates than other teens. Having a larger number of teenagers in the sample than would be expected proportionately in the population facilitates separate analyses of teenage sexual behavior, contraceptive use, and birth and pregnancy rates.

One of the goals of the 2006–2010 NSFG was to increase sample sizes

substantially, overall and for key subgroups. Sample sizes across subgroups did increase markedly from the 2002 NSFG to the 2006–2010 NSFG. Table H shows sample sizes by select subgroups for both surveys, and the absolute and percent increase from one survey to the next. Overall sample size increased 80%; all subgroups increased in size by at least 61%; and two subgroups, men and teenagers aged 15–19, more than doubled. As noted below in the ''Comparison of NSFG 2002 and 2006–2010 Standard Errors'' section, however, the *effective* sample sizes did not increase as much as the raw sample sizes did because of the variation in the weights and the clustering of the sample.

The growth of the Hispanic population made it critical for NSFG to obtain accurate information about the attitudes and behaviors of Hispanic men and women. It was important to have a Spanish version of the NSFG questionnaire as well. Using the Blaise system, interviewers could switch the language of the instrument with a single keystroke. The audio computer-assisted self-interviewing (ACASI) portion of the interview was also translated into Spanish. Table J presents the number of completed interviews with Hispanic men and women and the number of those conducted in Spanish. Of the 22,682 interviews conducted, 4,889 were with Hispanic respondents—3,308 in English and 1,577 in Spanish. About 32% of the interviews with Hispanic respondents, or 7% of all completed NSFG interviews, were conducted in Spanish (Table J). The probability of obtaining a representative sample of Hispanic respondents was greatly increased by interviewing in both languages. Potential respondents who did not speak English or Spanish were not included in the NSFG. In 2006–2008, this group accounted for about 0.6% of screened households and 0.5% of main interview households.

When seeking the voluntary cooperation of survey respondents, ISR sought to make the process as efficient as possible. An effective fieldwork process was vital to survey data quality. The steps involved in data collection and the letters, consent forms, and materials used in data collection are described in detail by Groves et al (1). In brief, these steps included sending an advance letter and brochure to all sampled households before contacting them in person to explain who was sponsoring the survey, who was conducting it, why it was being conducted, and that it was voluntary and confidential. The materials also cited toll-free phone numbers and the NSFG website as sources for additional information.

Interviewers then visited sample housing units, but in many cases they were unable to make contact with anyone in the sample household on a first visit. (Attempts to contact a sample unit are referred to as ''calls'' in the survey industry, but in the NSFG, ''calls'' refer to in-person visits.) Each time an interviewer visited a sample housing unit in person, she made a ''call'' (see Appendix I) to the housing unit. Interviewers often had to return several times until contact with the household was made. If contact was successfully made but the household member (or screener respondent) had no time to complete the screener, the interviewer and the household member tried to schedule a convenient later time.

When a field interviewer contacted a sample household, she introduced herself, displayed her identification badge, showed the authorization letter if necessary, and explained the purpose of the study. The interviewer then conducted a brief household screening interview to determine who in the household, if anyone, might be eligible for the NSFG. If there were no age-eligible persons (those aged 15–44) living in the household, no further contact with the household was made.

When a person aged 15–17 was selected for the sample, signed parental consent had to be obtained before the interviewer could talk to the selected minor. If the parent gave their consent, the minor was asked for his or her signed assent. If the parent or the minor did not give signed consent, the minor was excluded from the study. Emancipated minors—those aged 15–17 who were married or cohabiting and living away from their parents—were rare in a sample of this size, and when encountered, they were excluded from the sample.

Table H. Number of completed interviews in National Survey of Family Growth, 2002 and 2006–2010

Subgroup	2002	2006–2010	Sample size increase	Percent increase
Total	12,571	22,682	10,111	80
Sex				
Male	4,928	10,403	5,475	111
Female	7,643	12,279	4,636	61
Age				
15–19 years	2,271	4,662	2,391	105
20–44 years	10,300	18,020	7,720	75
Race and ethnicity				
Black	2,460	4,411	1,951	79
Hispanic	2,712	4,889	2,177	80
Other[1]	7,399	13,382	5,983	81

[1]Includes white, Asian, American Indian, and other races.

Table J. Number of Spanish-language interviews, by sex and race and ethnicity: National Survey of Family Growth, 2006–2010

Sex	All Hispanic persons	Hispanic, interview in English	Hispanic, interview in Spanish
Total	4,889	3,308	1,577
Male	2,297	1,570	723
Female	2,592	1,738	854

If the respondent was aged 18 or over, the interviewer gave the respondent an Adult Consent Form, which explained the survey and requested signed consent. If the adult agreed to participate but did not want to sign (this was very rare), the interviewer could sign for the respondent.

Two-phase sampling

The interviewer gave the respondent $40 as a token of appreciation for the respondent's help. The cost-effectiveness of these incentives (as they are called in the survey industry) had been established by experiments in previous cycles of NSFG (1,8,10). These experiments are also reviewed in Appendix II.

The token of appreciation was higher in weeks 11 and 12 of each 12-week quarter than in weeks 1–10 (phase 1). An experimental investigation conducted during quarters 2–4 indicated that an additional prepaid incentive of $40 mailed to households in phase 2 increased response rates and brought into the interviewed sample persons with characteristics that were of considerable interest to NSFG. The experiment in quarters 2–4 (see Appendix II for a detailed description) compared a $50 incentive in one-half of the phase 2 cases with an $80 incentive in the other half. The $80 incentive was more effective, bringing more distinctive respondents into the sample than the $50 incentive, including more Hispanic men, high-income men, childless women, and college-educated women. The 8% of respondents interviewed in phase 2 of the study (weeks 11–12) received $80. The 92% of respondents interviewed in phase 1 received $40.

It was important in the NSFG to monitor the number and types of calls interviewers were making to sample housing units because the number of visits is a major determinant of the cost of the survey; also, it was important to examine the final response rates. Table K shows the average number of calls, that is, in-person visits, to obtain a completed interview in the 2006–2010 NSFG. Visits to sampled households were recorded in the SurveyTrak sample management system on the interviewers' computers. Information about the type and time of the call as well as the outcome was recorded for each. On average there were 3.3 calls for every completed screening interview (Table K). Similarly, an average of 4.0 additional calls was needed to obtain a completed main interview once the screener interview was completed. Overall, approximately 7.2 calls were made per completed main interview.

The large number of in-person visits to obtain completed interviews is in part explained by field conditions accounted for in the sampling of housing units. A sample housing unit could be unoccupied when first visited. Or a housing unit might be occupied, but no age-eligible persons may live in the household. Several visits are often required before the interviewer learns whether anyone lives at the address, and whether there is anyone aged 15–44 in the household—that is, whether the address is occupied, and whether there is anyone in the household who is eligible for the interview.

More calls are often needed when access impediments are present. For example, sample housing units may be located in a locked building with no access to the housing unit from outside. Housing units may also be located in gated communities, where only residents and invited guests are permitted to enter.

Table L shows the rate of occupancy, impediments to access, and eligibility encountered in the 2006–2010 NSFG. About 86% of all sample housing units were occupied. Even unoccupied units may require more than one visit to verify occupancy status. Nearly 40% of housing units had access impediments such as locked apartment building doors and gated communities with guards. These impediments tended to require the interviewer to make more visits to determine if anyone lived there (i.e., to confirm occupancy). Among the occupied housing units (also known as households), only about 55% contained an age-eligible person. Interviewers visited many sample households repeatedly during data collection in order to determine whether anyone lived there, and if so, whether anyone was age-eligible.

After production interviewing for 10 weeks, households that had still not responded to the survey had been visited an average of 8 times. During weeks 11 and 12, phase 2 sample selection increased efforts for each remaining address, and (as described above) increased the incentive. (This approach was approved by the NCHS Institutional Review Board and the Office of Management and Budget.) This second-phase sampling sought as high a response rate as possible among housing units and screened households in the

Table K. Average number of calls (in-person visits) to obtain a screener, a main interview, and total to achieve main interview: National Survey of Family Growth, 2006–2010

Number of calls	Average (mean)
Number of screener calls to obtain screening interview	3.3
Number of main calls to obtain main interview[1]	4.0
Number of total calls to achieve main interview[2]	[3]7.2

[1]Mean number of calls per main interview is the average number of main calls on the cases with final main result code equal to 1001, a completed interview.
[2]Mean number of total calls on a case to achieve main interview is the average number of main and screener calls on the cases with final main result code equal to 1001, a completed interview.
[3]The average total calls is not equal to the sum of average screener and main calls due to rounding.

Table L. Percentage of occupied housing units, percentage of occupied housing units with access impediments, and percentage of occupied housing units with an age-eligible person: National Survey of Family Growth, 2006–2010

Occupancy status	Percent
All occupied housing units	86.2
Occupied housing units with access impediments	39.6
Occupied housing units with an age-eligible person aged 15–44	54.6

sample that were the most difficult to interview.

The second-phase sample also allowed examination of the potential size of nonresponse bias. Results from those interviewed in phase 1 were compared with those interviewed in phase 2 (some comparisons are shown in Appendix II). Lastly, the second-phase sample allowed the study to achieve a higher weighted response rate. Final response rates were higher in phase 2 than in phase 1 (2).

Paradata were used to select the phase 2 sample in each segment. Prediction models based on paradata generated daily predicted response propensities on the next day for all active addresses or sample persons. The prediction models were used to stratify segments and addresses remaining at the end of week 10 into two groups: those with a higher-predicted likelihood of interviews being completed and those with a lower-predicted likelihood of being interviewed. Within these two ''propensity'' groups, a sample of segments and then a sample of addresses within selected segments was chosen for phase 2 data collection.

Across all 16 quarters of data collection and among addresses still without final resolution at the end of week 10, 9,066 households had been successfully screened but not yet interviewed (Table B). Slightly less than one-third of the screened but not interviewed addresses (2,896 out of 9,066) were selected for phase 2 interviewing. Similarly, among the 8,082 addresses that had not been successfully screened for the presence of age-eligible persons by the end of week 10, a little less than one-third (2,368) were selected for phase 2 data collection.

The number of cases selected (or sampled) for phase 2 of fieldwork and the outcome of the phase 2 data collection are in shown in Table M (screener and main interview cases are shown separately). Main interview cases (those with a completed screener and an age-eligible person in the household) were selected at a higher rate than screener cases. Table M shows that 2,368 of the 5,264 cases selected for phase 2 were screener cases (45%) and 2,896 (55%) were main interview cases.

Among the 2,368 screener cases selected in phase 2, 932 yielded completed screeners in households with eligible persons, 758 yielded completed screeners in households with no eligible persons, and 678 were nonrespondent at the end of phase 2. Among the 2,896 main interview cases selected for phase 2, 1,757 yielded completed interviews, and 1,101 were nonrespondent at the end of phase 2. Among the phase 2 main interview cases, there were a small number of partial interviews, and 20 cases that were discovered to be nonsample addresses.

The 2,368 screener and 2,896 main interview phase 2 cases across all 16 quarters averaged about 329 per quarter—148 screener and 181 main interview cases, respectively. For an average quarter with 40 interviewers, the average interviewer workload for phase 2 consisted of eight (329 divided by 40) housing units—three screener interviews (148 divided by 40) and five main interviews (181 divided 40)—a substantial reduction in workload per interviewer sought in phase 2. Interviewers worked the same number of hours per week (approximately 30 hours) during phase 2 as in phase 1. Thus, an average of eight addresses per interviewer received approximately 60.0 hours of additional interviewer effort during phase 2, or about 7.5 additional hours per address.

The unweighted phase 2 screener response rate was not as high as the phase 1 rate. Among the 2,368 phase 2 selected screener cases, interviewers obtained 1,690 (932 + 758), or 71% completed screener interviews (Table M). The phase 2 main interview response occurred among those addresses initially selected from phase 1 (2,896) plus those found during phase 2 screening to contain an age-eligible person (932). Further, phase 2 review uncovered 20 main interview addresses that were found to be unoccupied. This left 3,808 main interview addresses eligible for interview in phase 2. The 1,757 completed main interviews in phase 2 came from both main interview and screener interview cases.

As in the 2002 NSFG (10), the total response rate and the phase 2 response rates are weighted. Respondents and nonrespondents are weighted by the inverse of the phase 2 subsampling probability. Thus, if the probability of selecting a case for phase 2 is one-third

Table M. Number of phase 2 screener and main-interview cases, by type of outcome, and response rates for phase 1 and phase 2 cases: National Survey of Family Growth, 2006–2010

Phase 2 (weeks 11–12) screener and main-interview cases	Number of cases
Screener	
Total cases	2,368
Completed screener, eligible	932
Nonrespondents	678
Completed screener, not eligible	758
Main interview	
Total	2,896
Completed interviews	1,757
Partial interviews	18
Nonrespondents	1,101
Nonsample	20

Phase 1 and phase 2 (weeks 1–12) main-interview cases	Response rate (percent)
Total	76.5
Phase 1 (weeks 1–10)	58.1
Phase 2 (weeks 11–12)	51.5

NOTES: Phase 1 response rates are unweighted; total, phase 2, and main response rates are weighted to account for phase 2 sampling rates. Phase 2 response rates are computed among eligible phase 2 cases. Weighted screener response rate was 93%.

(1/3), then that case has a weight of 3 times its base weight in phase 2. These phase 2 sampling rates varied across individual households, or main interview cases, with some rates receiving higher chances of selection and others receiving lower chances of selection. This variation in sampling rates required weighting all estimates, including response rates that included phase 2 cases.

The total response rate shown in Table M is weighted to account for the lower probability of selecting cases in phase 2 (2). The 77% rate consists of responses from phases 1 and 2, where the unweighted response rates were 58% and 52%, respectively (Table M). The 77% rate is also a combination of the screener and main interview response rates, 93% and 82%. These rates also follow standards described by Lepkowski et al. (2), in which addresses where eligibility had not been determined by the end of phase 2 were assigned as eligible or noneligible, so that those eligible cases could be included in the denominator of the response rate.

Table N shows the total sample and phases 1 and 2 response rates for key subgroups defined by sex, age group, and race and ethnicity. The total response rate is a weighted response rate, where housing units selected for phase 2 were weighted by the inverse of their phase 2 selection probabilities.

Overall, the final weighted response rate (shown in the fourth column of Table N) for the 16 quarters was 76.5%—77.7% for women and 75.1% for men. For male and female teenagers, the final response rate was 77.4% for female teens and 76.8% for male teens. The rate, however, did not show much variation across subgroups, from a low of 73.7% for Hispanic males to a high of 81.9% for black female teenagers. This was the result of responsive design features aimed at minimizing variation across subgroup response rates (see the next section for additional details).

As in Table M, there are two different types of response rates in Table N unweighted and weighted. The overall phase 1 rate of 58.1% is the unweighted percentage of eligible persons who, at the end of the first 10 weeks, had completed an interview. The overall phase 2 weighted response rate of 51.5% is the rate of completed interviews among known eligible persons achieved during the 2-week phase 2 period. This latter rate implies that 51.5% of the 41.9% of the phase 1 respondents, or 21.5% of the total sample of known eligible persons, were interviewed in phase 2. The combination of the phase 1 unweighted percentages and phase 2 weighted percentages, or 58.1% + 21.5% = 79.6%, might appear to be a suitable estimated overall rate. This simple sum does not capture the full effects of the weighting in the phase 2 rate. As a result, the final weighted response rate reported in the last column of Table N is slightly lower than the sum of unweighted rates, at 76.5%. Thus, Table N presents the difference between the unweighted phase 1 response rate and the final weighted rate—an 18.4% increase, not a 21.5% increase.

Table N also shows how the increase in response rates from phase 1 to phases 1 and 2 combined varied across subgroups of the sample. For example, the largest increase in the response rate in phase 2 (column 3) was for Hispanic males (22.2%).

Table N. Phase 1, phase 2, and final response rates, by sex, race and ethnicity, and age group: National Survey of Family Growth, 2006–2010

Sex and age group	Phase 1 (weeks 1–10) (1)	Phase 2 (weeks 11–12) (2)	Increase in response rate in phase 2 (3)	Final response rate (4) = (1) + (3)
		Percent		
Total	58.1	51.5	18.4	76.5
Female	59.4	52.7	18.3	77.7
Black	62.8	55.9	18.0	80.8
Hispanic	58.8	58.5	20.9	79.7
Other[1]	58.5	49.6	17.4	75.9
Ages 15–19 years	63.1	51.7	14.3	77.4
Black	65.8	62.3	16.1	81.9
Hispanic	62.4	50.3	13.8	76.2
Other[1]	62.5	49.4	13.9	76.4
Ages 20–44 years	58.6	52.9	19.1	77.7
Black	65.8	54.9	14.8	80.6
Hispanic	62.4	60.0	18.0	80.4
Other[1]	65.5	49.6	10.3	75.8
Male	56.6	50.1	18.5	75.1
Black	60.6	46.6	16.3	76.9
Hispanic	54.5	52.0	22.2	73.7
Other[1]	56.4	50.1	18.8	75.2
Ages 15–19 years	62.7	49.2	14.1	76.8
Black	66.1	39.6	10.6	76.7
Hispanic	63.1	52.6	13.4	76.5
Other[1]	61.5	50.5	15.5	77.0
Ages 20–44 years	55.0	50.3	19.7	74.7
Black	58.8	48.9	18.2	77.0
Hispanic	52.0	51.9	20.9	72.9
Other[1]	55.1	50.1	19.6	74.7

[1]Includes white, Asian, American Indian, and other races.

NOTES: Phase 1 response rates are unweighted; total and phase 2 response rates are weighted to account for phase 2 sampling rates. Phase 2 response rates are calculated as a percentage of those who have not been interviewed and are eligible for interview (ages 15–44) at the beginning of week 11 of interviewing.

Paradata and Survey Management

Paradata were used to make decisions about the survey design throughout the data collection process. The paradata used in NSFG included a set of observations that were monitored daily using the NSFG Dashboard (see Figure 9 of reference 1). The Dashboard tracked indicators of interviewer effort (such as hours of work or number of

visits to households) applied to the ''active sample,'' the housing units in the NSFG sample that were still being worked.

As with most household surveys, the active sample cases most easily contacted and most interested in participating in the study were interviewed earliest (11). As the first 10-week period (phase 1) proceeded in each quarter, the remaining active cases were those whose residents were rarely at home or whose lives left them only rare opportunities to participate in the survey. Efforts to obtain each interview increased over the days of the period. A key management challenge of the 2006–2010 NSFG was to direct those interviewer efforts over time to achieve a respondent pool that represented the full target population as much as possible within budget constraints.

As described above, study staff used paradata to generate two discrete-hazard- (logistic regression) response propensity models each night of the survey data collection throughout the 4 years. One model was for screener cases, predicting the likelihood that a screener interview would be completed on the next call. The second model was for main interview cases, predicting the probability of obtaining a completed interview on the next call.

Table O presents the coefficients obtained in the screener-response propensity model for the set of predictors found to be most highly associated with screener response. (Corresponding coefficients for the main-interview-response propensity model are not shown but are available on the NCHS website along with this report.) Predicted values, in the form of predicted probabilities for each case, were obtained from both the screener and the main model.

All of the coefficients shown in Table O are statistically significant in the total sample and for each quarter. This is to be expected given the large number of cases in the sample. The variable names shown in Table O are used in the subsequent descriptions of the outcome of the propensity model.

Many of the predictors used in the discrete-hazard-response propensity models were anticipated to be good predictors of daily screener or main interview outcomes. For example, there are known differences in response rates between urban and rural locations (12). The urban coefficient in the model (URBAN in Table O) indicates that those living in urban locations have lower predicted propensities to respond during the next visit. Similarly, the coefficients for evidence of non-English speakers in a segment (LNONENG), safety concerns expressed by the interviewer (LSAFECON), the presence of physical impediments to the entrance (PHYSIMPED), and addresses in multiple-unit housing structures (MANYUNITS) all indicate a lower propensity to respond during the next call.

Also, there are several operational measures that operate in an expected direction. A larger number of previous calls to an address (NUMPREVCALLS) indicates a lower propensity to respond. Ever having had a ''hard'' or a ''soft'' appointment with a household or person (PREVHARDAPPT or PREVSOFTAPPT) predicts a higher response propensity during the next call.

Table O also contains surprising associations. For example, negative comments and time-delay statements (PREVEVERSTATE) made by an informant predict a *higher* propensity to respond. Any informants at an address ever asking a question during prior calls (PREVEVERQUEST) reduced the propensity to respond, contrary to findings in the survey literature (12). The results in Table O are adjusted for a set of predictors that are different than those used in previous research.

The 2006–2010 NSFG had two principal responsive-design features that used these data: designing interventions during phase 1 to improve response rates in subgroups and selecting phase 2 cases. The predicted propensities were used to identify addresses where an interviewer might have a higher chance of obtaining an interview. In several interventions (see below), such cases were ''flagged'' for interviewer

Table O. Discrete hazard coefficients and odds ratios for daily screener response propensity models: National Survey of Family Growth, 2006–2010

Predictor name	Predictor description	Coefficient	Odds ratio
	Intercept	1.5499	
URBAN	Address in an urban location	–0.2738	0.760
LRESIDENTIAL	All housing units in sample segment are residential	–0.0115	0.989
LNONENG	Evidence of non-English speakers in sample segment	–0.2133	0.808
LSAFECON	Interviewer noted safety concerns about segment during segment listing or updating procedure	–0.0831	0.920
PHYSIMPED	Interviewer observed physical impediments to entry, such as locked door, community gate, etc.	–0.0254	0.975
MANYUNITS	Address in a structure with multiple housing units	–0.1253	0.882
NUMPREVCALLS	Number of calls made to this household prior to current call	–0.0140	0.986
PREVEVERCONTACT	Household ever been contacted	–0.4151	0.660
NUMPREVCONTACTS	Number of contacts made with this household prior to current call	–0.1635	0.849
PREVEVERSTATE	Informant at address ever made a negative statement or had time-delay in any previous call	–0.1789	0.836
PREVLASTSTATE	Informant at address made a negative statement or had time-delay in most recent call	0.1224	1.130
PREVEVERQUEST	Informant at address asked a question in any previous call	–0.0385	0.962
PREVLASTQUEST	Informant at address asked a question in most recent call	–0.0099	0.990
PREVRESISTDUMMY	Informant at address ever made statements indicating reluctance to be screened	–0.5570	0.573
PREVOTHCONTACT	Had contact with informant at address at most recent call	–0.1278	0.880
PREVSOFTAPPT	Ever had a soft appointment set with informant at address at any previous call	0.2995	1.349
PREVHARDAPPT	Ever had a hard appointment set with informant at address at any previous call	1.0828	2.953

attention, and there is some evidence that efficiency increased modestly.

The response propensities were used again at the end of week 10. The remaining nonrespondent cases were divided into screener and main interview subgroups. Within these subgroups there was a further subdivision into high-, medium-, and low-predicted propensity groups. A stratified random sample of nonrespondent cases was selected across these groups, or ''strata.'' Cases with higher propensity were given higher chances of being selected for the second phase than those in the medium- and low-propensity groups. The sampling rates, and whether there were three or only two propensity groups within screener and main cases, varied across quarters. Study staff sought to find a set of strata and sampling rates that maximized response rates and the number of completed interviews obtained in the second-phase sample, and did so with as few phase 2 calls as possible.

With the oversample of high-propensity strata in the phase 2 sample, the staff expected clear distinctions among high- and low-propensity stratum cases in production indicators, such as calls per case or calls to complete a screener or a main interview. Tables P and Q present a summary of production outcomes by data collection year for phase 2 cases. The purpose of examining the indicators summarized in these tables was to determine whether the propensity models and stratified selection led to higher response rates among high-propensity cases.

On average, it took fewer calls (visits) to obtain a screener or a main interview in phase 2 for high-propensity stratum cases than for low-propensity cases (see Table P). Among all screener cases, there was an average of 5.80 calls per high-propensity case, and 6.52 calls per low-propensity case. Thus, high-propensity cases required fewer calls, justifying an oversample to obtain a more efficient phase 2 sample. This pattern holds for both screener and main cases: high-propensity stratum cases required fewer calls than low-propensity stratum cases.

Table Q shows that the screener and main cases in high-propensity strata had higher response rates than those in the lower-propensity strata. For example, the screener response rate among selected phase 2 cases was 0.68 in the high-propensity stratum but 0.63 in the low-propensity stratum. Thus, oversampling high-propensity stratum cases generally led to slightly higher response rates in the phase 2 interviewing.

Table P. Average number of calls per selected phase 2 cases, by type of case and by phase 2 selection stratum (high- or low-predicted propensity of a completed interview on the next day): National Survey of Family Growth, 2006–2010

Type of case and propensity stratum	Weighted average, years 1–4
Screener	Calls per selected phase 2 cases
Total	6.16
High propensity	5.80
Low propensity	6.52
Main	
Total	7.59
High propensity	6.95
Low propensity	7.98

Table Q. Response rates for selected phase 2 cases and per completed phase 2 interview, by type of case and phase 2 selection stratum (high- or low-predicted propensity of a completed interview on the next day): National Survey of Family Growth, 2006–2010

Type of case and propensity stratum	Weighted average, years 1–4
Screener	Rates per selected phase 2 cases
Total	0.66
High propensity	0.68
Low propensity	0.63
Main	
Total	0.63
High propensity	0.65
Low propensity	0.62

Experimental Interventions in Fieldwork

The 2006–2010 NSFG also used interventions to address response rate and calling efficiency issues throughout the 16 data collection quarters (13). In each quarter at least one, and often several, interventions were developed to address nonresponse or other features of the design. Several interventions are summarized in this section.

Screener week

Every quarter had an intervention known as ''screener week'' (see Appendix I). Data from the 2002 NSFG suggested that some interviewers had a tendency to set aside screener cases in favor of main interviews (14). The result was that some interviewers had sample addresses that had fewer screener visits than other interviewers at the end of the data collection period. Propensity models suggested that some of these addresses had higher chances of yielding a completed screener compared with other addresses.

The 2006–2010 NSFG design had a much shorter data collection period than the 2002 NSFG—12 weeks compared with 11 months. Study staff were concerned that some interviewers might have large numbers of unsuccessfully screened cases at the end of the data collection period because they had not devoted enough effort to obtaining a completed screener interview.

Study staff decided to create a ''screener week'' starting in quarter 1, near the middle of phase 1. During screener week, interviewers were

encouraged explicitly to complete at least one call to all sample housing units, and to call as many screener cases as possible. Previously made appointments for interviews during the week were kept, but screener cases were emphasized during the week.

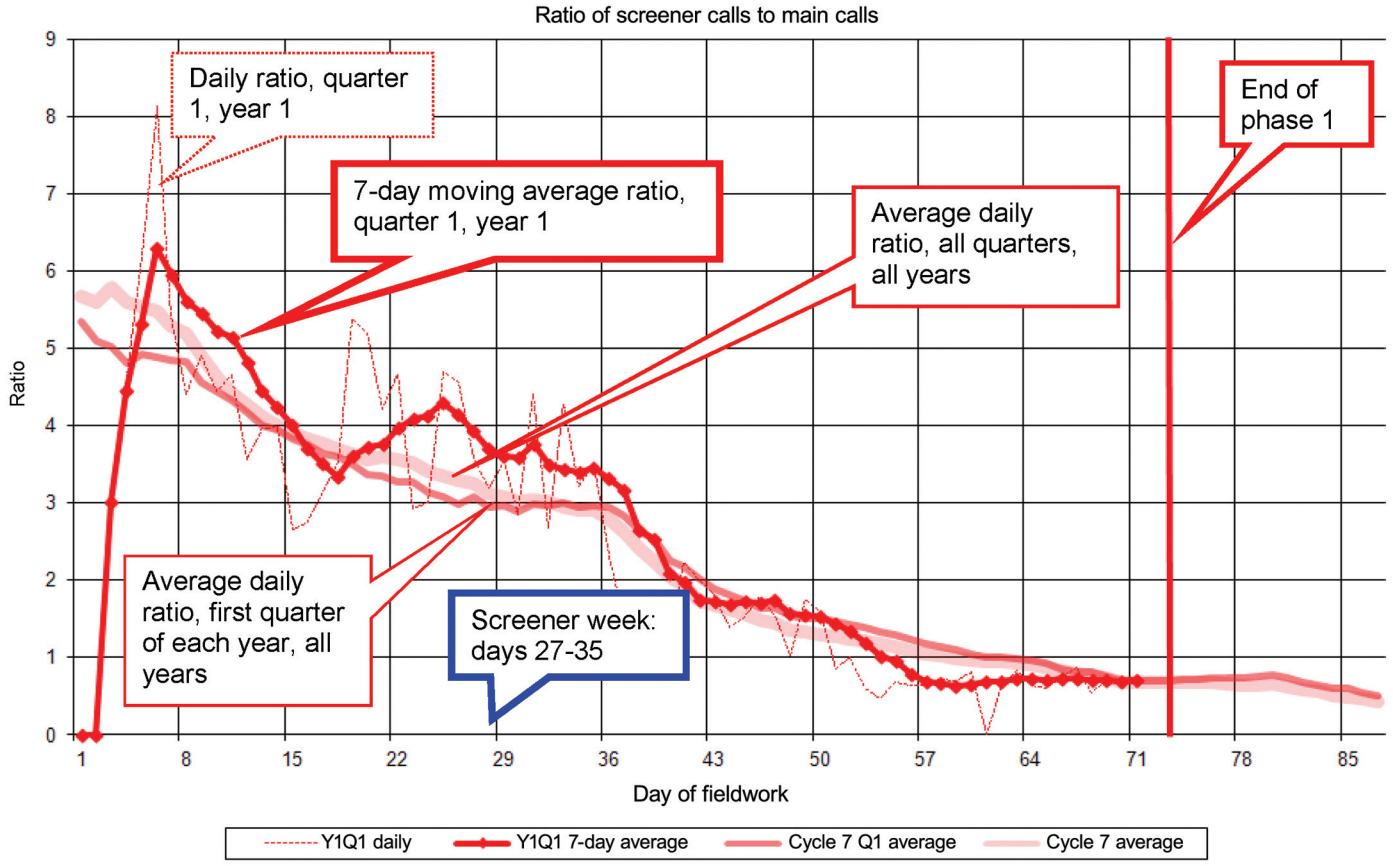
Screener week was typically in the fifth week of the 10 weeks in phase 1. The ratio of screener visits to main-interview visits typically declines rapidly as the quarter proceeds, but during screener week, that decline typically stops or slows down, because the number of calls for screeners increases and the number of calls for main interviews decreases (see Figure 2). In other words, screener week was successful in most quarters in that it increased calls to previously unscreened cases.

But considering another viewpoint, this effort was also aimed at increasing response rates. Despite the increase in the number of calls to screener cases during screener week, response rates increased only in one-half of the quarters. It is not yet clear under what circumstances screener week increases response rates; research on this issue is ongoing.

Other interventions

Several other types of interventions were implemented across the 16 quarters of data collection in NSFG. These interventions were reported and analyzed in a recent paper by Wagner et al. (13). For example, cases with low-base probabilities of selection, and therefore larger weights in analysis, were "flagged" in the interviewer address list in the sample management system, indicating that additional calls should be made to these addresses. If response rates were lower for groups with large weights, then larger nonresponse adjustments would be needed in the weighting process, and the variation in the weights would increase. Study staff tried to equalize the response rates for these under-sampled groups to prevent excessive weight variation in response rates across groups. Interviewers received messages from their supervisors urging them to increase calls to these marked units.

Many of these interventions were randomized at an interviewer level, where one-half of each interviewer's assignment was assigned to the intervention protocol, and one-half was assigned to a control group. Increased calls among the intervention cases occurred in nearly every such intervention. However, few interventions aimed at cases with large weights or high propensities to be interviewed led



NOTES: Y1 is year 1. Q1 is quarter 1.

Figure 2. Ratio of screener calls to main interview calls, by day of interviewing, showing the effect of "screener week" in the 2006–2010 National Survey of Family Growth

to higher response rates for the intervention group.

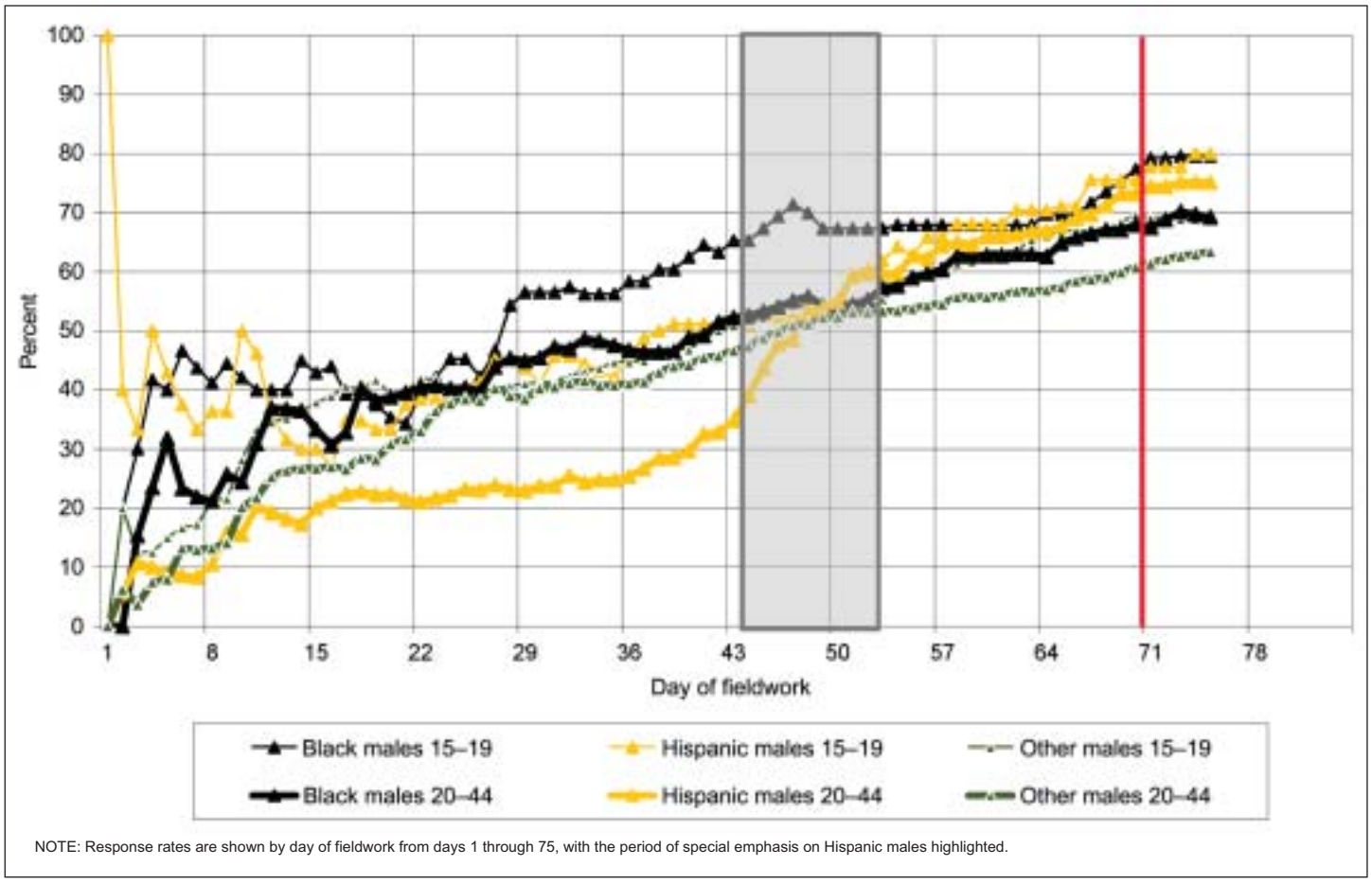
Another type of intervention was used to achieve improved sample balance (see Appendix I), and to reduce the variation in nonresponse adjustment weights. Key subgroups defined by sex, age, and race and ethnicity were monitored to see which groups had lower response rates compared with other subgroups.

Through the first year of data collection, response rates for Hispanic men aged 20–44 were consistently below all other key subgroups (see Figure 3). In quarter 5 the response rates were again monitored, and by week 6 it was clear that the response rate for Hispanic men aged 20–44 was again going to fall below the response rates of all other key subgroups. These Hispanic men (i.e., those identified in screener interviews) were flagged in interviewer sample management lists. Interviewers were encouraged to make more visits to these flagged cases. The result was a more rapid increase in response rates for this subgroup in the latter weeks of phase 1, as shown in Figure 3. That is, this intervention led to increased response rates among adult Hispanic men.

The effect of these NSFG interventions is difficult to evaluate overall. Some were randomized trials and showed increases in calls but showed little or no increase in response rates. Others were nonrandomized trials that appeared to have the desired effect on calls, response rates, or other outcomes.

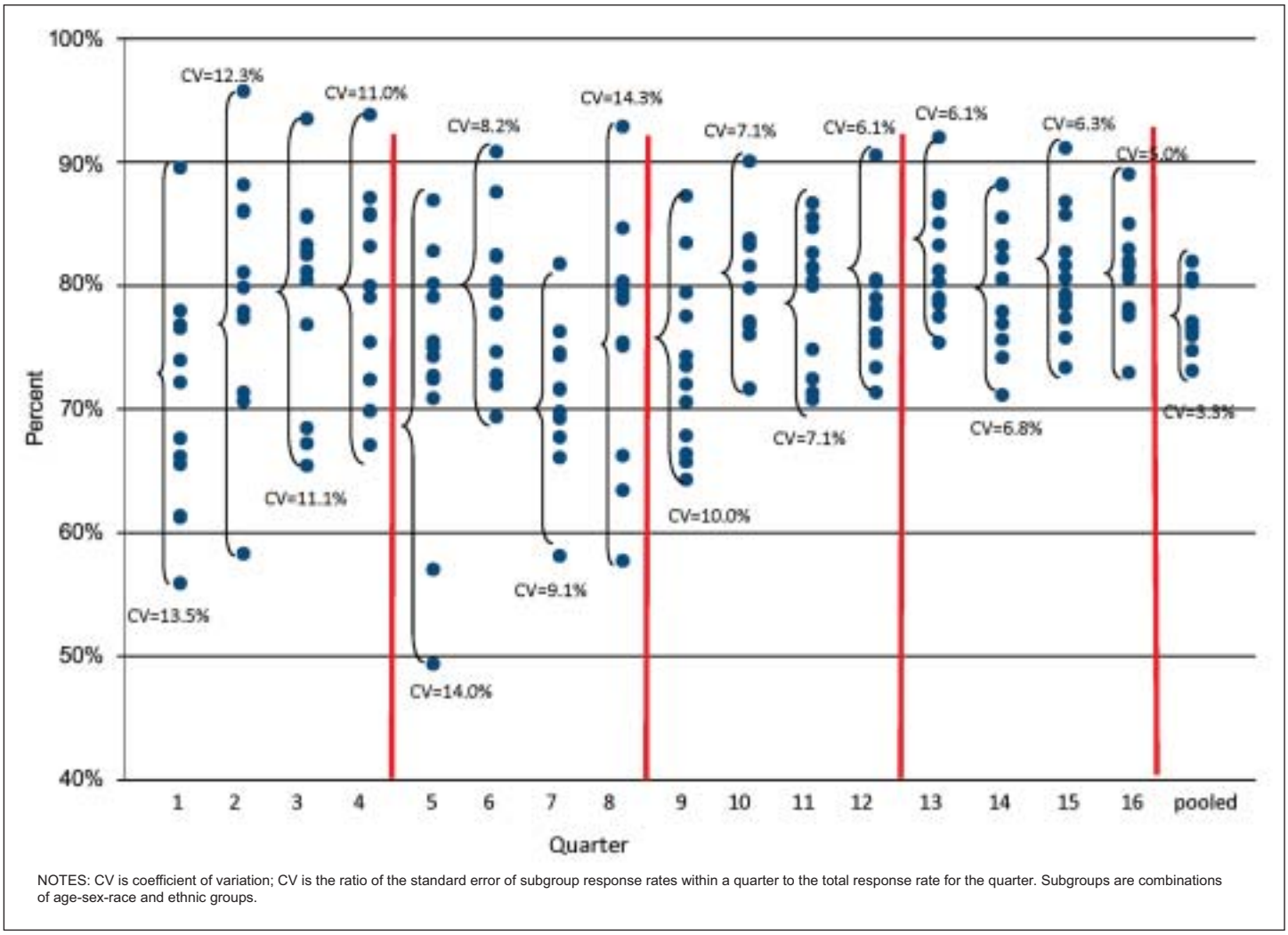
Study staff used another evaluation measure over the course of the 16 quarters: whether the interventions were having a beneficial effect on survey response rates. The response rates for 12 age-sex-race-ethnicity subgroups for each month were plotted on a graph, and the coefficient of variation of the response rates was computed. The coefficient of variation (CV) is a measure used in surveys to assess the extent to which variation in an outcome variable (in this case, response rates across 12 subgroups) is related to the level of the phenomenon (in this case, the overall response rate). That is, the CV is the standard error of the response rate variation across subgroups divided by the overall response rate.

Figure 4 presents results across 16 quarters for the coefficient of variation for response rates for 12 subgroups. As time progressed and interventions became more effective, response rates across subgroups converged. This trend is evident in the decreasing value of the coefficient of variation. For example, in early quarters, there was as much as a



NOTE: Response rates are shown by day of fieldwork from days 1 through 75, with the period of special emphasis on Hispanic males highlighted.

Figure 3. Response rates for subgroups of males, by race and ethnicity and age group in quarter 5 of the 2006–2010 National Survey of Family Growth



NOTES: CV is coefficient of variation; CV is the ratio of the standard error of subgroup response rates within a quarter to the total response rate for the quarter. Subgroups are combinations of age-sex-race and ethnic groups.

Figure 4. Response rates and coefficients of variation of the response rates for 12 subgroups for each quarter and all 16 quarters in the 2006–2010 National Survey of Family Growth

40 percentage point difference in response rates from lowest to highest across the subgroups. But in the last quarters the variance was less than 20 percentage points. The lowest response rates in Figure 4 are among Hispanic men aged 20–44. The higher response rates were among black female teenagers and black female adults consistently across the 16 quarters (data not shown).

The reduction in response rate variation across subgroups had two principal advantages. Sample sizes quarter by quarter were more consistent for subgroups and overall—making it easier to achieve specified target sample sizes for key subgroups across the quarters. Additionally, the improved sample balance across key subgroups reduced variation in nonresponse adjustment weights, resulting in less variation in the final weights.

Response Indicators

Response to the NSFG is a complex phenomenon. At the same time, the survey literature has long recognized that the response rate, as calculated above and shown in Table N, is not an entirely satisfactory measure of the impact of nonresponse on survey estimates.

For instance, a response rate is a single indicator for an entire survey, across all variables, subgroups, and estimates. Further, some survey questions have item missing values in addition to missing values for persons who do not respond to the survey at all.

Survey methodologists have developed alternative nonresponse indicators that tell more about the natureof nonresponse. Several indicators have been applied to the 2006–2010 NSFG. One such indicator—fraction missing information—is reported here and has appeared in previous survey literature (15). The interested reader should examine this reference for a more complete description of the application of the paradata-based fraction missing information indicator applied by Wagner to the 2006–2010 NSFG.

This variable-specific indicator captures another dimension of the response problem. In the absence of data for a given variable due to unit or item nonresponse, many surveys, including

the 2006–2010 NSFG, use available data from respondents to predict the missing values. These predictions take several forms, including nonresponse adjustments to weights or imputation of item missing values (see ''Weighting and Imputation'').

These individual predictions are themselves subject to variability. For example, in imputation, which replaces missing values with predicted values, no one replacement value is completely accurate. Multiple predictions can be made for each missing value if the error in the underlying prediction model (such as a predicted residual in a linear regression) is taken into account. This prediction error can be reduced when more highly correlated predictors are used in the model, but it cannot be eliminated.

In multiple imputations, several predicted values are generated for each missing value and data sets with reported and imputed values are created for sets of predicted values. Estimates are then computed from each data set, and the variation of estimates is calculated by combining variability within a multiple-imputed data set with variability in estimates among data sets. The between-data set variability is due to prediction error, and its contribution to the total combined error is referred to as the ''fraction missing information.'' This indicator has been proposed as a measure of the reliability of the imputed values provided in a survey data set.

Wagner applied this indicator to the 2006–2010 NSFG for a selected set of variables. Wagner multiply–imputed missing values for all known eligible persons in each quarter, regardless of whether the missing values were due to unit or item nonresponse. This was done for the final quarterly data set, but also for daily data sets—the 84 daily data sets in a quarter that contained all known eligible persons and survey responses for those who responded as of each day. That is, all missing values for a selected variable on a given day were replaced by an imputed value. The prediction was repeated multiple times, and estimates such as means or proportions were computed for the selected variables for each of the imputed data sets. Predictions for each variable were unreliable for day 1 when only a small fraction of the sample had responded; and by the final day, when 75% to 80% of the known eligible subjects had responded, the predictions were more reliable. For each day and variable, the fraction missing information indicator, a value between zero and one, was computed.

Wagner then displayed the values of the fraction missing information indicator across days for each variable. These displays showed a decreasing trend in the indicator across days, reflecting more reliable estimates each day. Further, the daily fraction missing information indicator was compared with the nonresponse rate for each variable, a competing indicator of the impact of nonresponse on estimates. Although the nonresponse rate indicates the proportion of missing values for a variable, it says nothing about what might be known from underlying correlations about the missing values themselves. The nonresponse rate was expected to be higher than the fraction missing information. Across monitored variables, there was clear evidence that the amount of missing information was less than the simple nonresponse rate for an indicated variable.

Whether the fraction missing information indicator becomes an accepted standard to report nonresponse impact for survey data is an open question. It is only a matter of time, though, before this new indicator or other proposed indicators will appear in summary reports alongside response rates.

Responsive Design Summary

The 2006–2010 NSFG employed a variety of sample and survey management strategies based on principles of responsive design. The features of the survey design were altered, as the paradata revealed problems in the survey fieldwork. This report provides only a summary of features such as two-phase sampling for nonresponse, screener week, interventions to achieve better sample balance, and response indicator monitoring. The 2006–2010 NSFG has been a part of a larger development of these kinds of methods, as discussed elsewhere (13,16).

Weighting and Imputation

The 2006–2010 NSFG used a complex sample design with over- and undersampling of key subgroups. Weights were necessary to correct for the over- and undersampling when combining cases across subgroups, and to compensate for nonresponse and noncoverage (see Appendix I for ''Sampling weight'' definition). There was also a compensation for item missing data (see ''Item imputation'' in Appendix I) in the survey. Rates of item missing data were low, but a sequential regression imputation procedure was used to replace item missing data on recoded variables.

The 2006–2010 NSFG weighting and imputation methods are described in detail elsewhere (2). This report presents summaries of the results of the final weighting and imputation processes, including specific predictor variables used in nonresponse adjustment models and descriptive characteristics of the weights at several stages of development.

Weighting Procedures

The NSFG weighting process had four major stages. First, a base weight was calculated that was the inverse of the probability that the case was selected.

Second, a nonresponse adjustment was determined as the inverse of the predicted probabilities of response from a logistic regression of response status (those who responded compared with those who did not respond). Separate nonresponse logistic-regression propensity models were estimated for screener and main interview cases on a set of predictors largely chosen from among the paradata available for respondent and nonrespondent cases. The nonresponse adjustment was applied to the base weights to obtain a nonresponse adjusted weight.

Third, the nonresponse adjusted weights were adjusted to U.S. Census Bureau projections of the number of persons in age-sex-race-ethnicity subgroups. Data from the Department of Defense's Defense Manpower Data Center on the number of military personnel living off-base in those subgroups were added to the Census projection values.

Finally, the population control-adjusted weights were examined to determine if any were so large to require reduction, or trimming (see Appendix I), to reduce excess variation in the weights. Given the large variations in the weights in this study, more trimming was necessary than in previous NSFG cycles.

Weights were produced for five subsets of the 2006–2010 NSFG sample:

- A final trimmed weight for analysis of all 16 quarters of the 2006–2010 data (adjusted to June 30, 2008 population control totals).
- Quarters 5–16 (adjusted to June 30, 2009 population control totals), because some new questions were introduced starting in quarter 5.
- Quarters 1–8 (the first 2 years of interviewing, adjusted to June 30, 2007 population control totals).
- Quarters 9–16 (the last 2 years of interviewing, adjusted to June 30, 2009 population control totals) because new variables were introduced in data collection beginning quarter 9.
- Quarters 1–10, the quarters for which the first public-use data file was released.

Base weights

Table R presents summary measures for the base weights. The base weight is the probability of selection for that case. It adjusts for unequal probabilities (see "Epsem" in Appendix I) of selection at the PSU level, the sample segment (within-domain) level, the housing unit level, the person-within-household level, and for second-phase sample selection. There is considerable variation in the weights at aggregate levels across age and racial and ethnic groups, reflecting the significant oversampling across age and racial and ethnic groups, and variations in sampling rates to keep interviewer workloads efficient.

For example, the base weight for black male teenagers has a minimum value of 48.07 and a maximum value of 12,550.60. The highest weights reflect variation due to the fact that some black teenagers come from areas that were sampled at lower rates, and from selection within households with larger numbers of other eligible persons and phase 2 subsampling.

Across racial and gender groups, the variation is larger. Among males, the minimum base weight is 24.70 and the maximum base weight 694,836.91. The larger variation occurs because white and other-race males have larger weights than do black males due to the lower sampling rates for "other" (mostly white) races. Combining other-race and black respondents adds to the variation present in the weights.

The final column in Table R is a measure of the variation of the weights relative to the average weight, abbreviated here by a standard notation $(1 + L)$ (17). In particular, the factor $1 + L$ is a measure of the relative variance of the weights themselves, and is computed as

$$1 + L = n \frac{\left(\sum_{i=1}^{n} w_i^2\right)}{\left(\sum_{i=1}^{n} w_i\right)^2},$$

where w_i is the weight for the i-th case in the sample and n is the number of sample cases. In the survey literature this measure is either referred to as a function of the coefficient of variation of weights or the potential loss in precision due to weighting. It is a global measure (that is, not specific to any one variable) that assesses the extent to which the variability of an estimated mean or proportion might be increased because some cases get small weights and others get large weights. A value of 1.0 indicates no contribution to variability due to weighting. A value of 2.0 suggests that there is a potential for the variability of estimates to double due to the weights.

The base weights in Table R show a substantial potential impact of weights on the variance of estimates. Table R shows that if the weights were all the same, increases in variance up to 9.5

Table R. Mean, minimum, and maximum untrimmed base weights, and potential increases in variance due to weighting (1 + *L*), by sex, race and ethnicity, and age group: National Survey of Family Growth, 2006–2010

Characteristic	Sample size	Mean weight	Minimum weight	Maximum weight	Increase in variance (1 + *L*)
Male	10,403	3,503.21	24.70	694,836.91	8.10
Black	1,854	2,208.36	45.49	170,072.47	6.14
15–19 years	470	1,958.38	48.07	12,550.60	1.83
20–44 years	1,384	2,293.25	45.49	170,072.47	7.17
Hispanic	2,297	3,176.04	40.11	180,321.32	4.80
15–19 years	551	2,724.68	46.33	62,106.06	3.29
20–44 years	1,746	3,318.48	40.11	180,321.32	5.08
Other[1]	6,252	4,007.39	24.70	694,836.91	8.64
15–19 years	1,361	3,204.84	37.65	56,509.01	2.57
20–44 years	4,891	4,230.71	24.70	694,836.91	9.50
Female	12,279	3,146.72	24.70	141,003.21	3.82
Black	2,557	2,209.78	38.06	141,003.21	5.46
15–19 years	462	2,257.85	38.06	28,171.71	2.48
20–44 years	2,095	2,199.18	42.15	141,003.21	6.16
Hispanic	2,592	2,822.26	38.06	73,633.46	3.66
15–19 years	484	2,752.16	38.06	38,425.09	2.80
20–44 years	2,108	2,838.35	38.06	73,633.46	3.85
Other[1]	7,130	3,600.68	24.70	114,409.90	3.47
15–19 years	1,345	3,223.89	35.07	32,950.78	2.26
20–44 years	5,785	3,688.29	24.70	114,409.90	3.67

[1]Includes white, Asian, American Indian, and other races.

times could occur if no adjustment or trimming of extreme weight values is done. But these adjustments were made, and the variation shown in Table R was reduced by the processes of nonresponse adjustment, poststratification, and trimming (described below).

Nonresponse adjustment of the weights

The next step in weighting was a nonresponse adjustment process. This two-part process included the screener interview and the main interview (2). The screener nonresponse adjustment had available as potential predictors only those variables available for all cases, largely paradata. However, one predictor included was not part of the standard paradata—an interviewer observation, or prediction of whether there were children under age 15 years in the household. Interviewers made this observation *before* attempting the screener interview based on her observations about housing unit and household characteristics.

This measure is worth including, in part because the presence of children under age 15 in the household is strongly correlated with many of the key outcome variables in NSFG. West reviewed the properties and accuracy of this correlation and a second interviewer observation about the marital and cohabitation status of household members that may affect the variability of the nonresponse adjustments used in the 2006–2010 NSFG (17).

The main interview nonresponse adjustment could use the variables available for the screener nonresponse model, the data collected in the screener (such as age, sex, and race and ethnicity), and an interviewer observation about whether the selected respondent is likely to be married to or cohabiting with an opposite-sex partner.

Table S presents the predictors used in the screener nonresponse-propensity models that generated the screener nonresponse adjustments. (For the sake of brevity, the values of the more than 50 coefficients are not given in Table S. A supplementary table with the full model may be obtained by e-mailing nsfg@cdc.gov.) Not surprisingly, many of these predictors are the same ones used in the daily response-prediction models discussed above. For example, urban location, evidence of non-English speakers in the sample segment, and physical impediments to entry all were associated with screener response propensity. More prominent among the screener response-propensity predictors are summaries of data recorded by interviewers during each contact with the household. A particularly effective predictor was whether an informant ever asked a question during one or more of the contacts (PREVEVERQUEST in Table S).

Table T presents the predictors for the main interview response-propensity model that generated nonresponse adjustment factors for main interviews. Several variables that were used in the screener nonresponse model appear in the main interview model, such as urban location. More important are variables from the screener interview, including whether the selected person is a teenager, a man, or a white person. Age is highly associated with many NSFG variables, and its inclusion as a predictor in the main interview nonresponse model means that the model has a combination of predictors that are associated with response propensity and, are themselves associated with the outcome variables in the survey. (For the sake of brevity, the values of all the coefficients in the model are not given in Table T. A supplementary table with the full model, which runs several pages, may be obtained by e-mailing nsfg@cdc.gov.)

The screener and main interview response-propensity models were used to calculate predicted probabilities of response for respondents and nonrespondents at the screener and main interview stages. A nonresponse adjustment weight was then calculated as the inverse of the predicted probability for each sample address and each eligible person, whether successfully screened or interviewed or neither. The screener nonresponse-adjustment weight was multiplied by the main interview nonresponse-adjustment weight for all main interview respondents. This nonresponse-adjustment weight was greater than one for all respondents, and set equal to zero for nonrespondents. The nonresponse-adjustment weight (sometimes called a nonresponse-adjustment factor) was subsequently multiplied by the base weight for each sample respondent to

Table S. Screener response propensity predictors for nonresponse adjustment models: National Survey of Family Growth, 2006–2010

Predictor name	Predictor description
URBAN	Address in an urban location (yes/no)
LRESIDENTIAL	All housing units in sample segment are residential (yes/no)
LNONENG	Evidence of non-English speakers in sample segment (yes/no)
LSAFECON	Interviewer noted safety concerns about segment during segment listing or updating procedure (yes/no)
PHYSIMPED	Interviewer observed physical impediments to entry, such as locked door, community gate, etc. (yes/no)
MANYUNITS	Address in a structure with multiple housing units (yes/no)
NUMPREVCALLS	Number of calls made to this household prior to current call
PREVEVERCONTACT	Household ever been contacted (yes/no)
NUMPREVCONTACTS	Number of contacts made with this household prior to current call
PREVEVERSTATE	Informant at address ever made a negative statement or had time-delay in any previous call (yes/no)
PREVLASTSTATE	Informant at address made a negative statement or had time-delay in most recent call (yes/no)
PREVEVERQUEST	Informant at address asked a question in any previous call (yes/no)
PREVLASTQUEST	Informant at address asked a question in most recent call (yes/no)
PREVRESISTDUMMY	Informant at address ever made statements indicating reluctance to be screened (yes/no)
PREVOTHCONTACT	Had contact with informant at address at most recent call (yes/no)
PREVSOFTAPPT	Ever had a soft appointment set with Informant at address at any previous calls (yes/no)
PREVHARDAPPT	Ever had a hard appointment set with Informant at address at any previous call (yes/no)

Table T. Main-interview, nonresponse-propensity model predictors: National Survey of Family Growth, 2006–2010

Predictor name	Predictor description
URBAN	Address in an urban location (yes/no)
LRESIDENTIAL	All housing units in sample segment are residential (yes/no)
LNONENG	Evidence of non-English speakers in sample segment (yes/no)
LSPANISH	Evidence of Spanish speakers in sample segment (yes/no)
LSAFECON	Interviewer noted safety concerns about segment during segment listing or updating procedure (yes/no)
PREVEVERCONTACT	Household ever been contacted (yes/no)
PREVEVERRESIST	Respondent or informant ever made statements indicating reluctance to be interviewed (yes/no)
MANYUNITS	Address in a structure with multiple housing units (yes/no)
PHYSIMPED	Interviewer observed physical impediments to entry, such as locked door, community gate, etc. (yes/no)
SCR_TEEN	Screener interview data indicate selected respondent is teenager (yes/no)
SCR_SEX	Screener interview data indicate selected respondent is male (yes/no)
SCR_RACE	Screener interview data indicate selected respondent is white (yes/no)
SCR_LANG	Screener interview data indicate anticipated interview will be in Spanish (yes/no)
SCR_SINGLEHH	Screener interview data indicate single person household (yes/no)
PREVEVERSTATE	Respondent or informant ever made a negative statement or had time-delay in any previous call[1] (yes/no)
PREVLASTSTATE	Respondent or informant made a negative statement or had time-delay in most recent call[1] (yes/no)
PREVEVERQUEST	Respondent or informant asked a question in any previous call[1] (yes/no)
PREVLASTQUEST	Respondent or informant asked a question in most recent call[1] (yes/no)
NUMPREVCALLS	Number of calls[1] made to this household prior to current call
NUMPREVCONTACTS	Number of contacts[1] made with this household prior to current call
PREVMAXRESISTDUMMY	Maximum previous resistance level
CHILD_LT15	Whether anyone in household is under 15 years of age based on housing unit observation (yes/no)
SEXUALLY_ACTIVE	Interviewer assessment after screener of whether respondent is married to or cohabiting with an opposite sex partner (yes/no)

[1]See Appendix I for definitions of calls and contacts.

obtain a nonresponse-adjusted weight before trimming.

Weight trimming

Table U presents summary measures for one part of the nonresponse adjustment factors before they were applied to the base weights for the main interview. There are large ratios of largest to smallest weights within many subgroups, but these large ratios do not have the same impact as for the base weights (presented in Table Q). This is because the nonresponse-adjustment factors are mostly between 1 and 2, with only a small number of cases with very large adjustment factor values. Thus, the $1 + L$ factors are, at most, 1.23 for Hispanic teen males.

Many survey organizations trim these types of adjustment factors separately from the other weights. These weights were trimmed in NSFG before proceeding to the population control adjustment, ''capping'' the nonresponse adjustment to a largest value of 4.0. The cap reduced weight variation somewhat, but given potential increases in variance in the untrimmed weight that were no more than a 26% increase in variance, the effect on the $1 + L$ factor was small. Only the untrimmed nonresponse adjustment factors for the main interview are shown here.

Finally, Table V presents summary statistics for the final weight—the nonresponse-adjusted base weight adjusted to population control values with excessively large values trimmed back to smaller values.

These final weights have been adjusted to amount to the population control totals after trimming. The mean weight for the 10,403 men is 5,972.08. Totaled among the 10,403 men, the final mean weight equals 62,127,548, which is the projected number of men aged 15–44 in the U.S. household population (including men in the military who are not living on military bases).

A final trimming process was applied to the weights because even after trimming nonresponse-adjustment factors, there was large variation in the preliminary final weights. Most of this variation is due to large variation in the base weights (see reference 2 for details of the weight calculation). Some of the variation could be attributed to efficiency improvement techniques, such as varying sampling rates within PSUs to achieve more equal workloads. However, the variation due to efficiency measures used in the survey was not the major source of variation in the weights.

The trimming process took the largest weights within each of the 12 age-sex-race-ethnicity subgroups of interest and reduced their value to the next largest value of the weights. These trimmed weights were then readjusted to the population control distributions. The number of weights trimmed within a subgroup varied by only 1%–2% to as much as 4% of the cases within the subgroup.

The trimming process deliberately reduces the variation in weights. The overall potential increase in variance due to weighting factor $1 + L$ for men is 2.41 for the 16-quarter weight, and 2.21 for women. Although still large, these factor values indicate that the final weight variation has been substantially reduced through the trimming process (compare, for example, 8.10 in Table R to 2.41 in Table V for males). These reductions in weight variation lead to smaller standard errors for estimates computed from the weighted 2006–2010 NSFG data.

On the other hand, trimming may also have negative effects on estimates. Trimming may change estimates substantially, particularly if the value of a variable with a large weight value is itself large. The trimming process included another step to assess whether large changes in estimates might be occurring as trimmed weights were created. Trimming occurred in a series of rounds within each subgroup. Each time the larger weights in a round were trimmed and readjusted to population controls, a set of 10 key male and 10 key female rates and means were

Table U. Mean, minimum, and maximum nonresponse adjustments and potential increases in variance due to adjustment (1 + *L*) for selected weight variable, by sex, race and ethnicity, and age group: National Survey of Family Growth, 2006–2010

Characteristic	Sample size	Mean weight	Minimum weight	Maximum weight	Increase in variance (1 + *L*)
		16 quarter weight			
Male	10,403	1.19	1.00	19.51	1.16
Black	1,854	1.16	1.01	5.48	1.11
15–19 years	470	1.15	1.01	4.32	1.11
20–44 years	1,384	1.16	1.01	5.48	1.11
Hispanic	2,297	1.21	1.00	7.82	1.18
15–19 years	551	1.22	1.00	7.67	1.23
20–44 years	1,746	1.20	1.01	7.82	1.16
Other[1]	6,252	1.19	1.01	19.51	1.17
15–19 years	1,361	1.16	1.01	6.74	1.13
20–44 years	4,891	1.20	1.01	19.51	1.18
Female	12,279	1.15	1.00	8.31	1.11
Black	2,557	1.14	1.00	7.93	1.10
15–19 years	462	1.13	1.00	5.64	1.08
20–44 years	2,095	1.14	1.00	7.93	1.11
Hispanic	2,592	1.15	1.00	7.21	1.10
15–19 years	484	1.18	1.01	7.00	1.16
20–44 years	2,108	1.14	1.00	7.21	1.09
Other[1]	7,130	1.16	1.00	8.31	1.12
15–19 years	1,345	1.17	1.01	7.36	1.13
20–44 years	5,785	1.16	1.00	8.31	1.12

[1]Includes white, Asian, American Indian, and other races.

Table V. Mean, minimum, and maximum final weights (after poststratification to Census data and trimming), and potential increases in variance due to the weights (1 + *L*) for selected weight variable, by sex, race and ethnicity, and age group: National Survey of Family Growth, 2006–2010

Characteristic	Sample size	Mean weight	Minimum weight	Maximum weight	Increase in variance (1 + *L*)
		16 quarter weight			
Male	10,403	5,972.08	38.70	49,735.07	2.41
Black	1,854	4,176.91	81.11	31,433.44	1.96
15–19 years	470	3,590.32	84.63	14,956.41	1.67
20–44 years	1,384	4,376.11	81.11	31,433.44	2.02
Hispanic	2,297	4,914.29	59.89	31,433.44	2.32
15–19 years	551	3,400.38	59.89	19,515.24	1.86
20–44 years	1,746	5,392.05	63.51	31,433.44	2.30
Other[1]	6,252	6,893.07	38.70	49,735.07	2.36
15–19 years	1,361	5,350.71	65.22	19,653.59	1.85
20–44 years	4,891	7,322.25	38.70	49,735.07	2.40
Female	12,279	5,029.30	41.08	30,226.35	2.21
Black	2,557	3,520.32	64.98	25,089.38	2.04
15–19 years	462	3,740.00	65.82	19,680.83	1.70
20–44 years	2,095	3,471.87	64.98	25,089.38	2.13
Hispanic	2,592	3,814.75	50.25	21,049.91	2.16
15–19 years	484	3,601.53	51.76	17,538.98	1.85
20–44 years	2,108	3,863.70	50.25	21,049.91	2.22
Other[1]	7,130	6,011.99	41.08	30,226.35	2.09
15–19 years	1,345	5,293.49	54.01	30,226.35	1.86
20–44 years	5,785	6,179.04	41.08	30,226.35	2.13

[1]Includes white, Asian, American Indian, and other races.

computed for the overall sample. If the change in estimates was more than 5% of the estimate value before any trimming began, the trimming step was not used. This limit was reached for only four key rates and means for trimming within two subgroups. The trimming process thus sought to reduce unnecessary weight variation while avoiding large changes in key survey estimates.

Discussion of weighting

The large variation in base weights is attributable to the over- and undersampling techniques in the stages of selection. For example:

- The weights for housing units across sample domains could vary by as much as 2.6 to 1 due to the oversampling of domains 2, 3, and 4.
- Weights also could vary by another factor of approximately 2 to 1 to adjust interviewer workloads to meet PSU-specific occupancy, eligibility, and response rates and differences in efficiency (i.e., the hours per completed interview).
- The housing unit weights varied across phase 2 strata, where some housing units were selected at rates 4 times larger than others in the second-phase sample. Thus, housing unit weights could vary by a factor of as much as 2.6 x 2 x 4 = 20.8 to 1.

All of these oversampling processes were designed to increase the efficiency of the data collection operations, and thereby reduce the per-unit cost of data collection. The purpose of oversampling domains 2, 3, and 4 was to obtain larger numbers of housing units in samples that had black or Hispanic persons. The oversampling to equalize workloads occurred because interviewers worked a fixed number of hours per week throughout each 12-week quarter—a management efficiency initiative. And the second-phase sample oversampling was implemented to yield a larger number of cases with higher propensities to respond to the phase 2 interview efforts.

Additional weight variation is attributable to the goal of increasing the number of teenagers aged 15–19 and the number of women in the sample. Within households, the sampling rates varied across teenagers, men, and women, with substantially higher rates assigned to teenagers and somewhat higher rates assigned to women. Black and Hispanic

teenagers were oversampled at the highest rates and other-race men aged 20–44 (primarily white adult men) had the lowest sampling rates, and consequently, the largest within-household selection weights.

The within-household variation in sampling rates was slightly different across black, Hispanic, and other-race households (see Figure 1). As the figure illustrates, the within-household probability of selecting a man aged 42 is 0.53/2.63, or 1 in 4.96. If this man had been selected in this household, he would have had a within-household probability of selection weight of nearly 5.

The illustration in Figure 1 is for a black household with one teenager aged 15–19. Consider a second illustration, an other-race household with two teenagers aged 15–19 and two adults (male and female) aged 20–44. Using the measures of size (see Appendix I) in Figure 1, the probability of selection of the older man is .022/2.16, or 1 in 9.82. If the older man had been selected from such a household, his within-household weight contribution would have been potentially 10 times larger than a man aged 20–44 living alone, who would have been selected with a probability of 1. Combining the maximum values of these oversampling weighting factors could account for a variation in base weights, in rare cases, of 20.8 x 10—or persons in the sample with base weights up to more than 200 times larger than other persons.

Finally, nonresponse adjustment and poststratification adjustment factors also contribute to weight variation. The nonresponse adjustment factors were deliberately trimmed to have no more than a four to one variation in values.

Returning to Table V, the variation shown in weights within various subgroups reflects the contributions of several factors:

- Oversampling subpopulations such as teenagers and black and Hispanic persons
- Equalizing interviewer workloads (which also causes variation in the weights and causes some neighborhoods to be oversampled)
- Oversampling in phase 2, the last 2 weeks of each 12-week quarter
- Oversampling teenagers within households
- Nonresponse adjustment
- Poststratification adjustment

In this situation, large weight variations are inevitable, and even the largest weights were simply products of very rare circumstances. For example, the largest preliminary final weights (prior to trimming) were other-race (usually white) men aged 20–44 living in households with teenagers in neighborhoods with high proportions of black and Hispanic persons (i.e., domains 2, 3, or 4); in areas with low occupancy, low eligibility, or low response rates; and large nonresponse-adjustment and poststratification-adjustment factors. Similarly, the smallest weight values were among black or Hispanic teenagers living in mostly white neighborhoods (i.e., domain 1); in areas with high occupancy, high eligibility, and low response rates with lower hours per interview; and small nonresponse-adjustment and poststratification-adjustment factors.

The trimming process was guided by two criteria: to reduce variation in weights and to leave weighted estimates unaffected. The first goal was monitored by the calculation of the potential increase in variance due to weighting factor $1 + L$. Weights were not trimmed if the impact on $1 + L$ overall or for a key sex-age-race-ethnicity subgroup was small (only a small percent change). Low weight values trimmed upward to higher weight values did not have any impact on the $1 + L$ factor. Trimming larger values had substantial impact.

The second criterion was to maintain the value of weighted estimates even after trimming. After each trimming operation, the weighted means of 10 key male and 10 key female variables were compared with those means computed with the preliminary final weight. Large changes (more than 5% in relative value) in the means were an indication that the weight trimming was changing weights for cases that had a substantial influence on the final weighted estimates. The trimming levels were no more than 5% of the cases within any one of the 12 key sex-age-race-ethnicity subgroups. Even at the maximum number of trimmed values in a subgroup, the final weighted estimates were very similar to the preliminary estimates computed using untrimmed weights. The relative difference approached 5% in only 2 of the 20 variables. In the remaining 18 variables the relative change was less than 1%.

Results of Imputation

The imputation process to replace item missing values in more than 650 recode variables in the final data set is, as noted above, described in detail in reference 2. Here, a summary of the amount of imputation is presented by type (sequential regression imputation and logical imputation; see "Item imputation" in Appendix I) for a set of frequently used female, male, and pregnancy recode variables.

Table W presents the number of reported, regression imputed, and logically imputed values for 23 selected recodes from the NSFG data files. For 20 out of the 23 variables shown in Table W, the percentage imputed is less than 1%. The two variables having the highest rates of item missing data and regression imputed values are measures of income: POVERTY (the ratio of household income to the poverty level) and TOTINCR (total income of the respondent's family). Because POVERTY is derived from TOTINCR, the imputation counts are identical for both variables. It is not unusual to find as much as 10% of values for family income and related income variables to be missing in survey data sets, so even though the missing data rate for these variables is higher than rates for other variables in NSFG, this occurrence is not unexpected.

The remaining variables show small and almost negligible levels of imputation, and represent what is present among the more than 650 recodes in the data sets. Very few cases and variables received logical imputation. The overwhelming majority of imputations were done by regression methods.

Table W. Sample size, regression imputed count, logically imputed count, and percent imputed for 23 selected recode variables: National Survey of Family Growth, 2006–2010

Variable name	Description	Sample size	Regression imputed	Logically imputed	Percent imputed
	Female and male variables				
ADDEXP	Additional births expected	22,682	64	8	0.32
AGEMOMB1	Age of respondent's mother at her first birth	22,682	382	0	1.66
EDUCMOM	Education of R's mother	22,682	329	0	1.45
HIEDUC	Highest degree received	22,682	8	8	0.08
LABORFOR	Labor force status last week	22,682	7	0	0.03
PARAGE14	Presence of parents at age 14	22,682	6	0	0.03
POVERTY	Poverty level	22,682	2,456	0	10.83
TOTINCR	Total income of the household	22,682	2,456	0	10.83
RELIGION	Religious affiliation at interview	22,682	65	0	0.29
	Male variables				
CSPBIOKD	Number of biological children R has fathered with current spouse or partner	3,909	2	0	0.05
LSEXUSE1	Contraceptive method used at last sex	8,630	42	0	0.49
SEX1MTHD1	Contraceptive method used at first sex	8,630	64	0	0.74
TIMESCOH	Total number of cohabitations	10,403	14	0	0.13
WANTB01	Wantedness of first birth in the last 5 years	1,256	2	0	0.16
	Female variables				
AGEBABY1	Age of woman at her first birth	12,274	3	2	0.04
CONSTAT1	Current contraceptive status	12,279	6	22	0.23
FECUND	Fecundity status	12,279	0	0	0
INFERT	Infertility status	5,422	4	17	0.38
MARDIS01	Date first marriage was dissolved	12,279	25	4	0.23
	Pregnancy variables				
AGEPREG	Age of woman at pregnancy outcome	20,492	84	5	0.43
DATEND	Date of pregnancy outcome	20,492	84	45	0.63
OUTCOME	Outcome of pregnancy	20,492	12	0	0.06
WANTRESP	Wantedness of pregnancy (respondent)	20,492	25	4	0.14

For each of the 23 variables in Table W and the remaining recodes in the female, male, and pregnancy data files, there is a corresponding imputation flag variable. The flag variable indicates which cases have been imputed logically and by regression, allowing the user to replace NSFG's imputed values with others generated by an analyst.

Variance Estimation

Estimates from the 2006–2010 NSFG are based on the sample of 22,682 respondents, rather than a complete enumeration of the eligible population of more than 120 million men and women aged 15–44. Consequently, the estimates are subject to error, a difference between the true population value and the value estimated from the sample. This difference may be due to systematic or fixed sources of error, such as nonresponse or noncoverage bias or due to variable sources of error, including the use of a sample to represent the population.

Probability sampling allows for the direct estimation of 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 reference 18 for a review of variance estimation techniques.

This section discusses the estimation of sampling variance (also called sampling error) for 2006–2010 NSFG estimates that accurately account for the principal effects of the different sampling techniques employed in the sample. There are three principal design features to account for: stratification of PSUs, selection of PSUs (cluster sample selection), and weights.

As discussed above, the 2006–2010 NSFG was based on a national multi-stage area probability sample. The entire land area of the 50 U.S. states and DC had known nonzero probabilities of being selected for this sample. As a result, every noninstitutionalized person in the United States had some probability of being selected for the NSFG sample.

In order to increase the number of interviews with black and Hispanic persons, geographic areas (neighborhoods) that have higher proportions of households with black and Hispanic persons are sampled at higher rates than areas with lower proportions. This sample design allows users to estimate characteristics of these important subgroups.

Variance estimation procedures are implemented in a number of computer software packages that are either commercially available or available for free online download. Some software packages allow users to estimate variances for means, proportions, regression coefficients, logistic

regression coefficients, and other statistics. There are two basic types of software systems available for this purpose: standalone and integrated packages. The standalone software packages require users to input data into a special format used by the system. Integrated software for estimation from complex sample survey data allows a user to conduct an analysis without converting data to another format. Once a survey data set is in a format used by a statistical software package such as SAS or Stata, estimates and test statistics within those systems that account for complex design features can be applied directly to 2006–2010 NSFG data.

NSFG 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 http://www.fas.harvard.edu/~stats/survey-soft/survey-soft.html (19). The site is maintained by the Survey Research Methods Section of the American Statistical Association. Heeringa et al. also provide guidance on how to use these software systems (20).

Using Survey Estimation Software

To illustrate the convenience of using these software systems for NSFG data, Figures 5–8 show how one basic analysis can be implemented using SAS, Stata, and SUDAAN. Additional illustrations of the use of these software systems for data analysis from the 2002 NSFG are available (21); (see also http://www.cdc.gov/nchs/nsfg/nsfg_cycle6.htm).

Figure 5 shows the estimation of percentages of women aged 15–44 who have ever had sexual intercourse and have ever been married, using the full 2006–2010 sample. The illustration includes the code and associated output for obtaining these estimates in SAS, version 9.2, using the PROC SURVEYFREQ command. The input

Program Code:

```
proc surveyfreq data = c7females;
    weight wgtq1q16;
    stratum sest;
    cluster secu;
    tables hadsex evrmarry / cl;
run;
```

Program Output:

```
                              The SURVEYFREQ Procedure
                                    Data Summary

                          Number of Strata                   56
                          Number of Clusters                152
                          Number of Observations          12279
                          Sum of Weights             61754741.1

              Whether Respondent has ever had sexual intercourse with a male (RECODE)

                             Weighted     Std Dev of          Std Err of 95% Confidence Limits
      HADSEX   Frequency    Frequency     Wgt Freq   Percent  Percent     for Percent
 ------------------------------------------------------------------------------------------
           1       10605     53475357      1662478   86.5931  0.6976    85.2083     87.9779
           2        1674      8279384       537887   13.4069  0.6976    12.0221     14.7917

       Total       12279     61754741      1936743   100.000
 ------------------------------------------------------------------------------------------

                                  Whether Respondent was ever married

                             Weighted     Std Dev of          Std Err of 95% Confidence Limits
    EVRMARRY   Frequency    Frequency     Wgt Freq   Percent  Percent          for Percent
 ------------------------------------------------------------------------------------------
           0        6745     28850952       941508   46.7186  0.9175     44.8974     48.5398
           1        5534     32903789      1294136   53.2814  0.9175     51.4602     55.1026

       Total       12279     61754741      1936743   100.000
 ------------------------------------------------------------------------------------------
```

Figure 5. Program output estimating proportions for two categorical variables for the female sample, SAS, version 9.2, in the 2006–2010 National Survey of Family Growth

data set is derived from the 2006–2010 NSFG female respondent file. The three key design features are specified in the WEIGHT, STRATUM, and CLUSTER statements, and the percentages are generated in the TABLES statement. In this case, based on the full survey period, the WEIGHT statement specifies the final 16 quarter weight (WGTQ1Q16). See the SAS documentation for other features on the SURVEYFREQ command (http://support.sas.com/documentation).

The program output shown in Figure 5 includes the count of strata (56), clusters (152), and the sum of the weights, which indicates the population size represented by these estimates. The output also provides the unweighted and weighted frequency counts and weighted standard deviations. The estimated percentages for both variables are also shown, along with the standard errors (computed using Taylor Series Linearization) for the estimated percentages and their 95% confidence limits.

The corresponding code and output for the same analysis using Stata, version 11+, are shown in Figure 6. A 'svyset' statement defines the variables that contain the sampling weights, strata, and PSUs. These values stay in effect until they are cleared or reset. The 'svyset' statement precedes the 'svy' command, which is used to analyze survey data. The output shown in Figure 6 repeats many of the values presented in the SAS output in Figure 5, and proportions rather than percentages are given in the table. The expression 'Linearized Std. Err.' refers to Stata's use (by default) of Taylor Series Linearization for computing variances for nonlinear statistics.

Finally, the illustration for the same estimation problem using the standalone SUDAAN software (version 10.1) is shown in Figures 7 and 8. The Figure 7 illustration processes an SPSS format version of the 2006–2010 female data set (filetype = SPSS), but other filetypes can be used as well. SUDAAN uses a very similar syntax to that in SAS, although the NEST statement is used instead of separate STRATUM and CLUSTER statements because SUDAAN has the capability to compute components of variance across multiple levels of a multistage sample. In addition, the data must be sorted in ascending order by the NEST variables before SUDAAN is able to process it for variance estimation purposes. (Note:

Program Code:

```
use "C:\c7females.dta", clear
svyset secu [pweight=wgtq1q16], strata(sest)
svy: prop hadsex evrmarry
```

Program Output:

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      56        Number of obs    =     12279
Number of PSUs   =     152        Population size  =  61754741
                                  Design df        =        96

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
hadsex       |
           1 |   .8659312   .0069764      .8520832    .8797792
           2 |   .1340688   .0069764      .1202208    .1479168
-------------+------------------------------------------------
evrmarry     |
           0 |    .467186   .0091751      .4489737    .4853984
           1 |    .532814   .0091751      .5146016    .5510263
--------------------------------------------------------------
```

Figure 6. Program output estimating proportions for two categorical variables for the female sample, Stata, version 11+, in the 2006–2010 National Survey of Family Growth

Program Code:

```
proc crosstab data = c7females filetype = spss design = wr;
     nest sest secu;
     weight wgtq1q16;
     class hadsex evrmarry;
     tables hadsex evrmarry;
```

Figure 7. Program code estimating proportions for two categorical variables for the female sample, SUDAAN version 10.1, in the 2006–2010 National Survey of Family Growth

Similar sorting is required for SAS and Stata variance estimation as well.) It is standard practice to specify in this kind of analysis that the PSUs were selected with replacement (design = wr), even when they are selected without replacement (for example, see reference 21).

The code for the two tables involves two statements, one to specify the grouping variables (the CLASS statement) and the other the table itself (the TABLES statement). SUDAAN's default output (Figure 8) is much more extensive than SAS or Stata for the same estimation. SUDAAN presents percentages and their standard errors, but it also presents estimated totals and their standard errors. The default variance estimation method in SUDAAN is the Taylor Series Linearization, as it was for SAS and Stata.

Why a Single-stage Variance Estimation Component is Sufficient

The 2006–2010 NSFG, like many area probability samples, has multiple stages of selection. These stages of selection each contribute to the variability of estimates computed from the sample. For example, consider an estimate from the 2006–2010 NSFG of the proportion of persons aged 15–44 who have ever been married. The estimated proportion is, as noted above, subject to variability due to sample selection. The sampling variance of the estimated proportion p, say $v(p)$, is a function of variability arising at the first and subsequent stages of selection in the sample. It is natural to expect then that when estimating the sampling variance for p it will involve components coming from the first and subsequent stages of selection.

Some NSFG analysts asked whether it is necessary to consider second and subsequent stages of selection in estimating the sampling variance of an NSFG estimate. This report offers a brief intuitive explanation (not a formal proof) here, and a more substantial demonstration in Appendix III, but it concludes that only the first-stage component is needed (for example see reference 22).

When computing $v(p)$ one must use estimates computed from a sample at the first stage. For the proportion of those who never married, the computation of $v(p)$ uses estimates of the proportion never married for each PSU, say p_α for the α^{th} PSU. But these PSU proportions are themselves based on samples involving the second and subsequent stages of selection. *The variability of p_α is incorporated into the first-stage variance component*. That is, each p_α variable brings into the first-stage component of variance variability due to the second and subsequent stages of selection. The first-stage variance component estimated using p_α automatically includes second and subsequent stages of variability. This variability due to second and subsequent stages is sufficiently accounted for by the single first-stage component.

Because only the first-stage component is needed, only first-stage units need be identified in the data set for use in variance estimation computations. Second- and subsequent-stage components do not need identification.

Analysts today raise the issue about second- and subsequent-stage components because of advances in the software used to compute variance estimates. Recent versions of publicly available software have an option to include identifying variables for second- and subsequent-stages of selection. This kind of feature is made available to survey statisticians wanting to estimate the variance components for each stage separately.

Estimates of the components are useful in the design of new surveys, but the estimated components are not needed for computing an estimate of sampling variance for an estimated proportion or other statistic. Appendix III provides an empirical demonstration of this result: the sampling variance is computed for several estimates using only the first-stage identification variable and using the first- and second-stage identification variables. The estimated sampling variances for the NSFG estimates are the same whether only first-stage or first- and second-stage identification variables are used. That is, Appendix III results show that only the first-stage component of variability is needed for NSFG analysts estimating sampling variance of NSFG estimates.

Variance Estimation Method: Taylor Series (WR)
by: whether respondent has ever had sexual intercourse with a male (recode).

		whether respondent has ever had sexual intercourse with a male		
		Total	1	2
	Sample Size	12279	10605	1674
	Row Percent	100.00	86.59	13.41
	SE Row Percent	0.00	0.70	0.70
	Lower 95% Limit ROWPER	.	85.15	12.08
	Upper 95% Limit ROWPER	.	87.92	14.85
	Col Percent	100.00	86.59	13.41
	SE Col Percent	0.00	0.70	0.70
	Lower 95% Limit COLPER	.	85.15	12.08
	Upper 95% Limit COLPER	.	87.92	14.85
	Tot Percent	100.00	86.59	13.41
	SE Tot Percent	0.00	0.70	0.70
	Lower 95% Limit TOTPER	.	85.15	12.08
	Upper 95% Limit TOTPER	.	87.92	14.85

by: whether respondent was ever married.

		whether respondent was ever married		
		Total	0	1
	Sample Size	12279	6745	5534
	Row Percent	100.00	46.72	53.28
	SE Row Percent	0.00	0.92	0.92
	Lower 95% Limit ROWPER	.	44.90	51.46
	Upper 95% Limit ROWPER	.	48.54	55.10
	Col Percent	100.00	46.72	53.28
	SE Col Percent	0.00	0.92	0.92
	Lower 95% Limit COLPER	.	44.90	51.46
	Upper 95% Limit COLPER	.	48.54	55.10
	Tot Percent	100.00	46.72	53.28
	SE Tot Percent	0.00	0.92	0.92
	Lower 95% Limit TOTPER	.	44.90	51.46
	Upper 95% Limit TOTPER	.	48.54	55.10

Figure 8. Program output estimating proportions for two categorical variables for the female sample (weighted totals and standard errors deleted), SUDAAN, version 10.1, in the 2006–2010 National Survey of Family Growth

As a result of the theoretical, intuitive, and empirical evidence, only the first-stage identification variable is included in the 2006–2010 NSFG public-use data file. This feature is consistent with previous NSFG cycles where only first-stage components are identified.

Comparison of NSFG 2002 and 2006–2010 Standard Errors

One might expect that because of the larger sample sizes available in the 2006–2010 NSFG compared with previous NSFG cycles, there should be substantial decreases in sampling variances and standard errors of estimates. In some cases the standard errors are much smaller in 2006–2010 than in 2002, but this is not always true. The standard errors of some statistics in NSFG may not be as small as might be expected with the larger sample sizes, because of design features that produced larger samples of teenagers and black and Hispanic persons at an affordable cost.

In complex samples like NSFG, several competing factors can affect the precision of estimates relative to previous cycles. One factor is sample size, which is *much larger in 2006–2010 than in 2002*, with increases of at least 60% for key sex-age-race-ethnicity subgroups (Table H). These sample size increases allow researchers using NSFG's much larger samples to analyze results for small subgroups of the population (teenagers, black and Hispanic persons, and adult men) more than ever before. These larger sample sizes would also be expected to increase precision.

A second factor, however, operates to decrease precision for the 2006–2010 NSFG compared with previous cycles. There is a larger *variation in the weights* in the 2006–2010 NSFG due to the goal of improving efficiency while increasing sample sizes of key subgroups. As noted above in the discussion of Table V, the larger weight variation can lead to increases in variances or to decreases in precision. The potential increase in variance due to weight variation factor $1 + L$ is an indicator that standard errors in the 2006–2010 NSFG may not be as small as the sample size increases would suggest. Further, the increase in variance is not expected to be uniform across subgroups. For example, the impact of the weight variation on black male teenagers aged 15–19 ($1 + L = 1.67$ in Table V) is far less than among all males ($1 + L = 2.41$). Estimates for ''all males'' are aggregated across groups that have substantially different probabilities of selection, such as teenagers and adults, and black persons and other race groups. The estimates for all males aged 15–44 are, therefore, affected by a larger variation in weights than in previous NSFG surveys.

The third factor affecting the precision of estimates in this sample is a *larger average cluster size* in the 2006–2010 NSFG than in 2002. The larger average cluster size occurs because the 2006–2010 NSFG has the largest sample sizes of any NSFG, and yet has about the same number of PSUs in the sample as most previous cycles. The average number of completed interviews per cluster is thus larger in the 2006–2010 survey. This is important because it is well known in survey sample design that larger average cluster sizes increase the effect of intra-cluster correlations of interviews. The ''design effect''—the ratio of (a) the sampling variance of an estimate (taking the cluster sample design into account) to (b) the sampling variance of a simple random sample (see Appendix I) of the same size—increases as the average cluster size increases.

A fourth factor must be considered also when comparing 2006–2010 NSFG estimated sampling variances with those from previous cycles. If estimates themselves are different, their standard errors will differ; an estimated proportion close to 0.5 will have a larger standard error than a proportion near 0 or 1. To control this effect, the examples shown here are for proportions that did not change much, so this factor is minor in the comparisons shown here.

In summary, the standard errors of some statistics in NSFG (primarily for adult white women) may not be as small as one might expect. The standard errors that did result, however, were the result of *more weight variation and larger average cluster sizes*, which were part of a successful effort to produce large samples of black, Hispanic, and teenage respondents at an affordable cost.

The 2006–2010 NSFG also has larger design effects, primarily because of the variation in weights and the larger average cluster size. The increase in design effects reflects a strategy for the 2006–2010 NSFG to stay within a fixed budget that was roughly equivalent to the annual budget for the 2002 survey; to reduce the chances of unexpected cost increases; to collect more interviews by reducing the cost per completed interview; and to allow for oversamples of black, Hispanic, and teenage respondents.

The design accomplished these goals. For example, the 2006–2010 sample was 80% larger than in 2002—22,682 interviews in 2006–2010 compared with 12,571 in 2002. This increase in sample size was achieved in part by a cost per case that was about 35% lower in 2006–2010 than in 2002. And the response rate was 78% for women and 75% for men, about the same as in the 2002 survey. This increase in sample size makes possible many analyses that were not possible in the 2002 and prior NSFGs because adequate case counts were not available.

Table X shows several examples from the female, male, pregnancy, and teenage files. The examples include overall samples as well as subgroups determined by Hispanic origin, race, and age. These examples were chosen because oversampling occurred by sex, age, and race and ethnicity. Sampling variances for estimates that combine age or racial groups will tend to have larger weight variation and potentially smaller gains in precision, despite sample size increases.

A total of 19 comparisons of standard errors are shown in Table X, including data from the male, female, and pregnancy files. *On average, there was a mean reduction in standard errors (a gain in precision) of 11.1% for the 19 estimates in Table X from 2002 to 2006–2010; the median change was a 12.0% reduction in standard errors from 2002 to 2006–2010 for these 19 estimates.* The largest reduction in

Table X. Estimated standard errors for estimated percentages in four subgroups, by race and ethnicity, age group, and sex: National Survey of Family Growth, 2002 and 2006–2010

Subgroup	2002			2006–2010			Standard error percent change from 2002 to 2006–2010
	n	Estimated percent	Standard error	*n*	Estimated percent	Standard error	
Percentage of contraceptors who were using the oral contraceptive pill							
All	4,619	30.6	0.93	7,304	27.5	1.02	9.7
Hispanic	921	22.0	1.40	1,568	19.8	1.62	15.7
Non-Hispanic white	2,546	34.4	1.17	3,891	32.0	1.43	22.2
Non-Hispanic black	853	22.7	1.92	1,325	18.3	1.41	–26.6
Non-Hispanic other	299	25.4	2.62	520	20.6	2.53	–3.4
Percentage of men who intend to have a(nother) birth							
All, 15–44 years	4,928	55.4	1.22	10,403	58.9	1.00	–19.0
15–19 years	1,121	89.5	1.23	2,378	92.1	0.79	–35.8
20–24 years	938	85.0	1.50	1,733	86.7	1.37	–8.7
25–29 years	708	71.8	2.32	1,807	73.7	1.64	–29.3
30–34 years	724	47.6	2.59	1,555	51.8	2.03	–21.6
35–39 years	746	29.1	2.25	1,500	31.3	1.98	–12.0
40–44 years	691	16.5	1.68	1,430	15.7	1.37	–18.5
Percentage of women and men aged 15–19 who have ever had sexual intercourse							
Women	1,123	45.5	1.80	2,255	42.6	1.70	–5.6
Men	1,112	45.7	2.10	2,371	41.8	1.60	–23.8
Percentage of single live births in the last 5 years that were breastfed at all							
All	2,270	67.5	1.69	4,499	69.3	1.72	1.0
Hispanic	745	74.7	2.39	1,224	75.0	1.84	–2.3
Non-Hispanic white	1,299	69.4	2.37	1,896	72.9	2.33	–1.7
Non-Hispanic black	540	47.8	2.90	1,046	46.0	2.63	–9.3
Non-Hispanic other	186	68.4	6.24	333	72.5	4.97	–20.4

standard errors shown in Table X was 35.8% (for male teenagers—the percentage who intend to have a future birth), while the largest increase in standard errors was 23.0% (for breastfeeding among births to Hispanic women).

In summary, the gains in precision from the 2002 survey to the 2006–2010 survey were greatest in the *oversampled groups*—Black, Hispanic, and teenage respondents. In general, standard errors for comparable statistics were smaller consistently in 2006–2010 for men in all age and racial groups; smaller in 2006–2010 for black women; and remained about the same or were slightly larger in 2006–2010 for white women and women of all races.

For white females and all females, sample sizes increased but standard errors did not always shrink from 2002 to 2006–2010 because of the large increases in average cluster size and variation in weights due to oversampling.

Conclusion

This report presents an overview of the responsive design approaches to the design and fieldwork of the 2006–2010 NSFG; the outcomes of fieldwork; and weighting, imputation, and variance estimation procedures. NSFG accomplished its goals of increasing sample sizes substantially for key subgroups, reducing the cost per interview, and achieving response rates exceeding 75%. In order to accomplish these goals, it was necessary to introduce a variety of new responsive design procedures, including:

- Collecting and analyzing paradata during 16 quarters in a 4-year period to make continuous improvements in data collection.
- Monitoring response rates, the sample yield of key subgroups, and survey costs and productivity to produce a data set that was within budget and contained the needed oversamples.
- Analyzing the effects of the survey design (including weighting and clustering) on variance estimates for the survey and using trimming and other procedures to reduce weight variation to the extent possible.

Given the scope of these changes in how the survey was designed, two preliminary reports (1,2) provided advance notice of these changes, and this report provides the outcomes of these changes. This information should

be useful to NSFG analysts in preparing their own research and to survey methodologists who may wish to consider some of NSFG's methods for their own work.

References

1. Groves RM, Mosher WD, Lepkowski JM, Kirgis NG. Planning and development of the continuous National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 1(48). 2009.
2. Lepkowski JM, Mosher WD, Davis KE, et al. The 2006–2010 National Survey of Family Growth: Sample design and analysis of a continuous survey. National Center for Health Statistics. Vital Health Stat 2(150). 2010.
3. French DK. National Survey of Family Growth, cycle 1: Sample design, estimation procedures, and variance estimation. National Center for Health Statistics. Vital Health Stat 2(76). 1978.
4. Freedman R, Whelpton PK, Campbell AA. Family planning, sterility and population growth. New York: McGraw-Hill. 1959.
5. Whelpton PK, Campbell AA, Patterson JE. Fertility and family planning in the United States. Princeton, NJ: Princeton University Press. 1966.
6. Ryder NB, Westoff CF. Reproduction in the United States, 1965. Princeton, NJ: Princeton University Press. 1971.
7. Ryder NB, Westoff CF. The contraceptive revolution. Princeton, NJ: Princeton University Press. 1977.
8. Kelly JE, Mosher WD, Duffer AP, Kinsey SH. Plan and operation of the 1995 National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 1(36). 1997.
9. Groves RM, Heeringa SG. Responsive design for household surveys: Tools for actively controlling survey errors and costs. J Royal Stat Soc A 439–57. 2006.
10. Groves RM, Benson G, Mosher WD, et al. Plan and operation of cycle 6 of the National Survey of Family Growth. National Center for Health Statistics. Vital Health Stat 1(42). 2005.
11. Olson K, Groves RM. An examination of within-person variation in response propensity over the data collection field period. J Official Stat 28(1):29–51. 2012.
12. Groves RM, Couper MP. Nonresponse in household interview surveys. New York: John Wiley and Sons, Inc. 1998.
13. Wagner J, West BT, Kirgis N, Lepkowski JM, Axinn WG, Kruger-Ndiaye S. Use of paradata in a responsive design framework to manage a field data collection. J Official Stat 28(4):477–99. 2012.
14. Kirgis N, Lepkowski JM. Design and management strategies for paradata-driven responsive design: Illustrations from the 2006–2010 National Survey of Family Growth. In: Kreueter F (ed). Improving surveys with paradata: Analytic uses of process information. New York: John Wiley and Sons, Inc. 2013.
15. Wagner JT. The fraction of missing information as a tool for monitoring the quality of survey data. Public Opin Q 74(2):223–43. 2010.
16. Kreueter F. Improving surveys with paradata: Analytic uses of survey process information. New York: John Wiley and Sons, Inc. 2013.
17. West BT. An examination of the quality and utility of interviewer observations in the National Survey of Family Growth. J Royal Stat Soc A 211–25. 2013.
18. Wolter KM. Introduction to variance estimation, second edition. New York: Springer Science + Business Media, LLC. 2007.
19. Section on Survey Research Methods, American Statistical Association. Summary of survey analysis software. Available from: http://www.hcp.med.harvard.edu/statistics/survey-soft/.
20. Heeringa SG, West BT, Berglund PA. Applied survey data analysis. Boca Raton, FL: Taylor and Francis Group, LLC. 2010.
21. 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.
22. Cochran WG. Sampling techniques, 3rd edition. New York: John Wiley and Sons, Inc. 1977.
23. Mosher WD, Pratt WF, Duffer AP. CAPI, event histories, and incentives in the NSFG Cycle 5 pretest. In: 1994 Proceedings of the American Statistical Association, Survey Research Methods Section. Alexandria, VA: American Statistical Association. 1994.

Appendix I. Glossary

Audio computer-assisted self-interviewing (ACASI)—Interview method in which the respondent uses a laptop computer to complete a questionnaire. The interviewer asks the respondent to use earphones that deliver an audio recording of the questions. The question text is also displayed on the laptop monitor for the respondent to read. The respondent clicks or keys in the answer to each question, using the laptop mouse or keyboard. The software directs the respondent to the next appropriate question based on the answers entered. In the 2006–2010 National Survey of Family Growth (NSFG), the respondent performed these steps out of the sight of the interviewer to offer the respondent as much privacy as possible. ACASI was offered in both English and Spanish in the 2006–2010 NSFG.

Blaise—Software system that presents the interview questions in a questionnaire such as NSFG. Blaise is programmed to route the respondent to the next appropriate question, store the respondent's answers, and check the consistency of one answer with answers to other related questions. Blaise was used in the 1995, 2002, and 2006–2010 NSFGs.

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

Computer-assisted personal interviewing (CAPI)—Interview method in which the interviewer uses a laptop computer. The laptop displays question text for the interviewer to read, and provides any other necessary instructions to the interviewer. Interviewers record the respondent's answers using the keyboard. Software directs the interviewer to the next appropriate question based on the answers entered.

Contact, contact rate—Interviewer visits to a household in the sample during which the interviewer speaks with someone who lives there. The contact rate is the percentage of sample households where an interviewer talked with someone at the household at the screener stage (i.e., the screener contact rate). At the main interview stage, the contact rate is the percentage of those selected for interview who actually met with the interviewer during one or more visits to the household by the interviewer (i.e., the main interview contact rate).

Delivery sequence file (DSF)—U.S. Postal Service listing of all addresses to which mail is currently delivered by the Postal Service. In most areas, the DSF is the basis for a list of housing units from which listings for NSFG are created.

Design effect—Ratio of the actual variance for a statistic to the variance in a simple random sample of the same size. The design effect, averaged over some representative statistics from a survey, gives an indication of the effect of the complex design of the survey on the precision of the statistics from the sample.

Domain—Stratum, or group of sampling units (such as blocks), placed in the same subset from which a sample of units was selected.

Double (or two-phase) sample—Subsample of nonrespondent sample cases, selected after the completion of a phase of data collection. NSFG used such a subsample in Cycle 6 (2002) and in 2006–2010.

Eligible household—Household containing at least one person who is eligible for NSFG—that is, persons aged 15–44 at the date on which the screener was completed and living in the household population of the United States (all 50 states and the District of Columbia). It is unknown whether a selected household has an eligible person until the household screener is conducted. If a household has two or more persons aged 15–44, one person is selected randomly.

Eligibility rate—Percentage of sample cases that are members of the target population. In NSFG, the eligibility rate is the percentage of households that contain a person aged 15–44.

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

Institute for Social Research (ISR)—University of Michigan organization that conducted the fieldwork and data processing for the 2006–2010 NSFG under a contract with the National Center for Health Statistics (NCHS). ISR has several centers that participate in NSFG: the Survey Research Center provides overall coordination and is responsible for data collection, weighting, and variance estimation; the Inter-university Consortium for Political and Social Research processes data and develops documentation and Web-based systems; and the Population Studies Center provides substantive expertise on demography and family growth.

Institutional Review Board (IRB)—Committee of peer and community reviewers of research procedures involving human subjects that weighs the benefits of the research relative to the risks of harm to human subjects. NSFG was reviewed and approved by the NCHS IRB, which NCHS calls a Research Ethics Review Board (RERB).

Intervention—In the 2006–2010 NSFG, a number of changes in interviewing practice that were made during interviewing to resolve imbalances in the sample or to address problems that arose during fieldwork. The most commonly used intervention was ''screener week,'' during which uncompleted screener cases were flagged to prompt interviewers to give those cases priority until the screener was completed. Other interventions were

''weight-based,'' meaning that cases with large weights (e.g., adult males) were flagged, and interviewers were asked to give those cases priority. ''Propensity-based'' interventions were those in which cases with a high probability of resulting in an interview were given priority for interviewer work.

Item imputation—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 650 recoded variables, or ''recodes'' (defined below), rather than on all of the thousands of variables in the data set. The purposes of imputation are to make the data more complete, more consistent, and easier to use, and to reduce bias caused by differential failure to respond. For example, if a respondent's educational level is missing and a value of ''high school graduate'' is assigned, education is imputed. Imputation was done in two ways in the 2006–2010 NSFG: logical and regression imputation. Regression imputation uses a regression equation to estimate a value for a case with missing data. Regression imputation was used to assign most of the imputed values. Occasionally, however, logical imputation was used: Logical imputation uses a subject-matter expert to assign a value based on the value of other variables for the case with missing data. For nearly all recoded variables in the 2006–2010 NSFG, less than 2% of the cases received an imputed value.

Main interview—Interview sought within sample households containing a person aged 15–44. If the screening interview shows that the household contains one or more persons aged 15–44, a main interview is requested from one of those persons selected at random.

Measure of size or size value—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 primary sampling units (PSUs) are the count of occupied housing units obtained in the 2000 Census of Population and Housing, because sample selection within the PSU would have selected housing units. Measures of size are also used in the selection of eligible persons within the household (see Figure 1) to increase the chances of selection of such groups as teenagers (aged 15–19), black and Hispanic persons, and females. Each person in the household is assigned a measure of size between zero and one, where the measures are predetermined values for each age by gender and racial and ethnic group. The measures of size are cumulated across eligible persons, a random number from zero to the sum of the measures is generated, and an individual is selected based on the cumulated measures of size.

Multi-phase design—Survey design that changes its sample design or recruitment protocol over different sets of sample cases or over time periods of the survey to obtain optimal balance of costs and quality of survey estimates. NSFG was a multi-phase design because it used two phases in each quarter.

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

Office of Management and Budget (OMB) clearance—OMB reviews survey materials and questionnaires proposed for use by government agencies under the provisions of the Paperwork Reduction Act. The review is conducted by the OMB's Office of Information and Regulatory Affairs.

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 used observations of characteristics of sample housing units to reduce the number of callbacks, used statements made by household screener informants in order to diagnose their concerns about the survey, and used call record data to model the probability of obtaining an interview on the next visit. Some paradata are labeled as ''process data.''

Phase—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 the 2006–2010 NSFG, there are two phases in each 12-week quarter. In weeks 1–10, the standard protocol is used, although paradata are used to optimize the efficiency of the interviewers. In weeks 11–12, a subsample of nonrespondents from phase 1 is offered higher incentives and certain other rules are changed (see ''Double sample'').

Primary sampling unit (PSU)—First-stage selection unit in a multistage area probability sample. In the NSFG, PSUs are counties or groups of counties in the United States; 110 PSUs were selected into the NSFG sample for 2006–2010.

Race and ethnicity—Three categories of race and ethnicity were used in the 2006–2010 NSFG for purposes of sample design: Hispanic, non-Hispanic black, and all other. Race and ethnicity is used in this report *as it was used to select the NSFG sample*. Hispanic and non-Hispanic black persons were selected at higher rates than others in the NSFG in order to obtain adequate numbers of Hispanic and black persons to make reliable national estimates for these groups. Thus, in this report, tables showing race and ethnicity show the three categories *used to design and select the sample*. However, in reports that are designed to present *substantive results*, OMB guidelines for the reporting of race are followed. For example, the ''all other'' category is often split into ''non-Hispanic white'' and ''non-Hispanic other'' categories; and respondents who report more than one race are classified separately from those who report only one race.

Recoded variables or recodes—Variables selected from the NSFG data

file that are to be constructed, edited, and imputed. NSFG staff selected about 650 variables because it is not possible to edit or impute all of the variables in the 2006–2010 NSFG data file. Recodes are variables that are likely to be used frequently by NCHS and other data users. They are edited for consistency, and missing values are imputed. Many (but not all) of these recoded variables are constructed from other variables in NSFG; some are constructed from a large number of other variables. Other variables in the data file are not edited or imputed in this way.

Replicate—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 in order to control the workflow of the interviewers. In responsive designs, early replicates are used to measure key cost and error features of a survey.

Respondent—Person selected into the sample who provides an interview. In the 2006–2010 NSFG, respondents were the approximately 5,500 men and women aged 15–44 who completed the NSFG interview each year.

Response rate—Respondents to a survey divided by the number of eligible persons in the sample. In this report, the response rate is the number of respondents (aged 15–44) divided by the number of eligible persons (aged 15–44). Given that not all screeners were completed, the number of eligible persons is not known precisely. The number of eligible persons is estimated among all screeners that were not successfully completed and added to the denominator of the response rate.

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; and combine data from the separate design phases into a single estimator. The 2002 and 2006–2010 NSFGs were both based on responsive designs.

Sample balance—A deliberate effort by study staff to equalize response rates across key sex-age-race-ethnicity subgroups. Sample balance was monitored by examining the distribution of response rates across 12 key subgroups and computing the coefficient of variation of response rates across subgroups. In early quarters, there was as much as a 40 percentage point difference in response rates from lowest to highest across the subgroups, compared with less than 20 percentage points in the last quarters. See Figure 4 and the accompanying text.

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

Sampling weight—Estimated number of persons in the target population that a respondent represents. For example, if a man in the sample represents 12,000 men in his age and race-ethnicity category, then his sampling weight is 12,000. The NSFG sampling weights adjust for different sampling rates (of the age and race-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 of these factors, it is sometimes called a ''fully adjusted'' sampling weight.

Screener week—One-week period during each 12-week sample quarter when interviewers were encouraged explicitly to complete at least one call to all sample housing units, and to call as many screener cases as possible. Appointments for interviews during the week were kept, but the emphasis was placed on screener cases during the week. Screener week typically occurred during the fifth of the 10 weeks in phase 1.

Screening interview—Set of questions (usually short) asked of a household informant aged 18 or over to determine whether the household contains anyone eligible for the survey. In the NSFG, the screening interview (sometimes called a ''screener'') consisted of a household roster collecting age, race and ethnicity, and sex. Households having one or more persons aged 15–44 were eligible for a main interview.

Segment—Group of housing units located near one another, all of which were selected into the sample.

Self-representing area—County or group of counties forming a PSU with population counts large enough to equal or exceed the typical stratum size in the U.S. national sample. Such PSUs are thus, always included in the sample. The sampling probabilities for persons in such areas are designed to be equal to sampling probabilities in smaller PSUs, called nonself-representing areas.

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

Strata or stratification—The partitioning of a population of sampling units into mutually exclusive categories (strata). Typically, stratification is used to increase the precision of survey estimates for subpopulations important to the survey's objectives. In the 2006–2010 NSFG, those groups include

teenagers (aged 15–19), 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. See reference 2 for further details.

SurveyTrak—Software-based sample administration system used in the 2006–2010 NSFG. The system is used by interviewers on laptop computers to document their sample assignment, organize the activities of their workday, remind them of appointments, record results of each call attempt, record observations of the sample housing unit, and track their job duties in all other ways.

Target population—Population to be described by estimates from the survey. The target population in NSFG was the household population of the United States, which refers to the civilian noninstitutionalized population and active-duty military personnel who are not living on military bases. Noninstitutionalized refers to the omission of persons residing in prisons, hospitals, dormitories, and other large residences under central control. College students living in dormitories were interviewed but sampled through their parents' or guardians' households.

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

Weight—See ''Sampling weight.''

Weight-based intervention—See ''Intervention.''

Appendix II. Research on Incentives Used in the National Survey of Family Growth: An Overview

The National Survey of Family Growth (NSFG) has a history of providing ''incentive'' payments to respondents, and of testing alternative levels of payment. (In the materials provided to sampled households, incentives are called ''tokens of appreciation,'' so this phrase or the more common ''incentives'' will both be used here.) Incentives in the NSFG take the form of cash at the time the interview begins. Four major experiments with incentives have been conducted in NSFG. This appendix describes the first three of these briefly, and then describes the most recent experiment in more detail.

1993 (Cycle 5) Pretest

In a field experiment in the 1993 pretest for NSFG Cycle 5, a $20 payment was found to produce a significantly higher response rate (67.4%) than when no payment was offered (58.9%) (8). For women who were offered $20, field costs per case were also lower than for women who received no incentive (23).

2001 (Cycle 6) Pretest

In a field experiment in the 2001 pretest, a $20 payment was contrasted with a $40 payment. The response rate for those offered $20 was 62%, and for those offered $40, it was 72%. Respondents receiving the higher amount were also less likely to express objections or reluctance to the interview than those receiving $20 (21).

Cycle 6 Main Study

In the 2002 Cycle 6 Main Study, a $40 incentive was used but response rates were still lagging in key groups after 7 months of interviewing. NSFG staff requested and received permission from the Office of Management and Budget (OMB) to use an $80 incentive in a half-sample of *remaining nonresponding cases* in the final 4 weeks of data collection during February 2003. The $80 incentive raised the weighted response rate from 64% to 79%. The sample in the last 4 weeks had a higher proportion of married women, Hispanic men and women, and full-time workers of both sexes (10).

These experiences showed cost-effective increases in response rates and sample representativeness with the use of incentives.

2006–2010 NSFG

The fieldwork for the 2006–2010 NSFG included obtaining a household roster to see if someone aged 15–44 lived in the household (the ''screener''), and obtaining a main study interview. Each data collection year was divided into four 12-week quarters. For 10 weeks, interviewers offered a $40 token of appreciation during attempts to contact households and to obtain interviews. The response rate by the end of the 10 weeks in year 1 of interviewing averaged about 58%—a rate that was determined to be too low by the National Center for Health Statistics (NCHS) and NSFG cosponsoring agencies. By the end of week 10, interviewers had visited nonresponding households an average of 8 times in person to obtain a screener and an interview. Given the cost of these visits in interviewer time and expenses, it was necessary to consider ways to improve the odds of success.

In week 11 of each quarterly data collection period (the beginning of phase 2), study staff drew a subsample of the remaining unfinished cases. The remaining cases were generally of two types: non-contacted cases in which the interviewer is unable to talk to the selected respondent (e.g., completing a screener with Mrs. Jones; Mr. Jones is selected as the respondent, and the interviewer is unable to find him at home for an interview); and time-delay statements in which the respondent is found at home, but says that he or she cannot give the interview at the present time (e.g., ''I can't do it this week because I'm sick,'' or ''We're having dinner now, come back later'').

Initially, NSFG sought to use the same procedure that was approved by OMB and the NCHS Institutional Review Board and that was used successfully in February 2003 in Cycle 6. Phase 2 cases were offered a higher token of appreciation—$80 instead of $40. The first $40 was prepaid and delivered in an overnight letter. The remaining $40 was paid when the respondent sat down to begin the main interview. In Cycle 6, 724 out of 12,571 respondents, or 6%, received the higher incentive. (Note that the $80 incentive is never offered to teenagers aged 15–17 but only to adults.)

These experiences showed cost-effective increases in response rates and in sample representativeness. The results led to a proposal to use a $40 incentive during phase 1 of the 2006–2010 NSFG, and an $80 incentive in the phase 2 nonresponse follow-up. NCHS management suggested conducting an experiment to determine whether it was necessary to increase the incentive from $40 to $80, or whether staff could increase it from $40 to $50 and obtain a similar improvement in response rates while reducing the overall cost of data collection. Subsequently, a randomized experiment was conducted during quarters 2, 3, and 4 in 2006 and 2007 to determine whether the increased incentive would increase response rates and yield a sample with better balance across key subgroups.

Design of the Experiment

The basic experimental design operated within the 2006–2010 NSFG 12-week quarter. During phase 1 (weeks 1–10 of each quarter), selected survey respondents aged 15–44 were offered $40 to complete an interview. During phase 2 (weeks 11 and 12 of each quarter), a sample of approximately one-third of the remaining cases was selected. Adults aged 18–44 in quarters

2, 3, and 4 were randomly divided into two groups:

a) Group 1 received $10 prepaid in addition to the standard $40 (a total of $50 for the main interview).

b) Group 2 received $40 prepaid in addition to the standard $40 at the beginning of the interview (a total of $80 for the main interview).

These two groups are designated as the $10/$40 and the $40/$40 experimental conditions. Cases selected for phase 2 were sent a final letter via express mail with the prepaid incentive enclosed. The letter stated that the enclosed incentive was for the respondent to keep in appreciation for their help. (For conciseness, the $40/$40 group is referred to as the "$80 group" and the $10/$40 group is referred to as the "$50 group.")

Table I below presents the pooled results across the three quarters. The table shows counts of cases and response rates separately for household screener and main interview cases in order to evaluate the potential impact of the incentives in each group. Overall, the response rates for screener cases in phase 2 that were offered $80 for the main interview were *10 percentage points higher* than for the screener cases offered $50 for the main interview (77% compared with 67%, respectively). The response rates for main interview cases in phase 2 were *12 percentage points higher* in the $80 group than the $50 group (64% compared with 52%, respectively).

NSFG study staff also wanted to know if the $80 incentive brought different types of people into the sample compared with the $40 or $50 incentives (Table II). The sample sizes in the experimental categories are small because only a one-third subsample of the remaining cases were selected for phase 2, and that subsample was randomly split into two payment plans (for women, n = 51 in the $50 group and n = 68 in the $80 group), but the sample size in the phase 1 group is 1,896. The split among men is similar: n = 47 in the $50 group and n = 70 in the $80 group; the sample size in the phase 1 group is 1,432.

Despite the small subsamples in the phase 2 groups, 9 of the 13 differences between the phase 1 $40 and $80 groups were significant using two-tailed t tests. Only 3 of the 13 comparisons between the $40 group and the $50 group were significant. These results suggest that *the $80 incentive was recruiting different people into the sample* compared with the $40 group, but the $50 incentive was much less effective in that respect. Given due caution about the sample sizes, the patterns are clear:

For women:

- Sixty percent of the $80 group was childless at the date of interview, compared with 41% of women who received $40. Because the principal outcome variable of NSFG is fertility and birth rates, this is a *critical finding: the $80 incentive was more effective in including childless women in the survey*. (There was no significant difference in the group that received a $50 incentive.)
- Among women who received $80, only 24% lived in multiunit structures (e.g., apartments and condominiums), compared with 38% of women who received $40. (This was also true for men: 26% compared with 37%, respectively.) These findings suggest that both men and women living in single-family homes were more likely to respond to the higher incentive.

For men:

- Hispanic men were a larger percentage of men in the $80 group (37% of men were Hispanic), compared with only 20% of the $40 group. No significant difference was observed between the $40 and $50

Table I. Pooled quarters 2, 3, and 4 unweighted case counts for phase 2 incentive experiment outcomes, response rates, and simple random standard errors: National Survey of Family Growth, 2006–2010

	Screener interview cases in phase 2						
Experiment	Sample size	Completed screener interviews	Refusal	Noninterview	Nonsample	Response rate (percent)	Standard error (percent)
$10/$40	208	130	43	20	15	67	3.4
$40/$40	207	152	32	14	9	77	3.0

	Main-interview cases in phase 2						
Experiment	Sample size	Completed main Interviews	Refusal	Noninterview	Nonsample	Response rate (percent)	Standard error (percent)
$10 prepaid	192	100	48	44	0	52	3.6
$40 prepaid	215	137	29	49	0	64	3.3

NOTES: Teenagers aged 15–17 were not included in the experiment. Their token of appreciation for the interview was never more than $40. Randomized assignment of phase 2 cases to treatment groups was made on the segment level (e.g., all cases in a segment were assigned to the same treatment group). Therefore, the simple random sample standard errors are likely to underestimate the true standard errors.

Table II. Comparison of sample characteristics between the $10/$40 and $40/$40 experimental groups in quarters 2, 3, and 4 in token of appreciation experiment: National Survey of Family Growth, 2006–2010

Characteristic	Phase 1 $40	Phase 2 $50	Phase 2 $80
Female			
Sample size	1,896	51	68
College degree or more	34	48	†51
Ever had an abortion	6	3	†1
Never had a live birth	41	68	†60
Ever had sex with a female	13	16	†4
Income $75,000 per year or more	17	†40	25
Living in a multi-unit structure	38	‡24	†24
Male			
Sample size	1,432	47	70
Hispanic	20	24	†37
College degree or more	28	‡43	36
Never fathered a birth	57	52	64
Ever had sex with a male	7	5	†1
Income $75,000 per year or more	25	3	†42
Living in a multi-unit structure	37	42	‡26

† Two-sided hypothesis test comparing phase 1 and phase 2 experimental groups statistically significant at the $\alpha = 0.05$ level.
‡ Two-sided hypothesis test comparing phase 1 and phase 2 experimental groups statistically significant at the $\alpha = 0.10$ level.

groups of men. Given strong policy and program interest in representing Hispanic men in national surveys, this is also a key finding.

- Among men in the $80 group, 1% had sex with another man compared with 7% of men in the $40 phase 1 group. Given the strong public health-related interests in these behaviors, this is also a critical finding.
- Among men who received $80, 42% had incomes of $75,000 or more a year, compared with 25% of men who received $40. This suggests that the $80 incentive was more effective in drawing high-income men into the sample.

Conclusions

Despite relatively small samples in the two experimental groups, consistent results were obtained across three consecutive quarters: the $80 incentive raised response rates and recruited different people into the sample than did the phase 1 effort alone ($40 incentive). Further, the results are broadly consistent with findings from the 2002 NSFG. The results suggest that *childless, high-income people living in single-family homes* are not as well represented in the standard phase 1 sample as they are when offered $80. Drawing these people into the sample improves the representativeness of the sample and raises the response rate. This appears to justify the use of the $80 incentive for a small subset of the sample.

Given the costs of conducting the experiment and its effects on response rates, NCHS and study staff asked permission to end the experiment after three quarters of data collection. NCHS' Institutional Review Board, or its Research Ethics Review Board, granted this request on August 29, 2007 (Amendment 11, NCHS Protocol Number 2006–01); OMB granted the request on November 21, 2007. The $80 incentive in phase 2 was adopted in all subsequent quarters of the 2006–2010 NSFG.

Appendix III. Accounting for Multistage Sample Designs in a Single-stage Variance Estimate

The 2006–2010 National Survey of Family Growth (NSFG), like many area probability samples, is based on multiple stages of selection. It is well known (22) that the sampling variance of the mean includes a component for each stage of selection. It would seem reasonable to assume then, that in order to estimate the sampling variance of an estimate or perhaps the standard error of a mean or proportion, the calculation must use information on all stages of selection in the calculation. That is, a complete survey data set for a multistage sample must have variables that would allow the identification for each case of the primary sampling unit (PSU), the second-stage unit, the third-stage unit, and so on.

Lesser-known is that to estimate the sampling variance of an estimate from a multistage survey, one does not need to calculate components of variance from each stage. When working with sample data, the sampling variance estimation only needs to use the first-stage units.

This result can be explained in terms of the theoretical properties of variance estimates in multistage surveys, but those theoretical treatments are not accessible to the everyday user of multistage sample survey data.

But empirical evidence exists that can be used to prove the result. This section examines the estimation of sampling variances and standard errors from the 2006–2010 NSFG that uses only the first stage or PSU level of clusters in the survey, as well as estimation using the first-, second-, third-, and fourth-stage components. The sampling variances of estimated means and proportions are compared across the methods, and are found to be quite similar. And because using only one stage in estimation is easier and somewhat faster, this section explains why NSFG surveys, including 2006–2010 and all previous cycles, provide a first-stage unit identifier on the file, and not second-, third-, or fourth-stage units.

Multistage Selection Compared With Single-stage Selection

Some survey analysis software systems allow a user to incorporate into the variance estimation calculation components for each stage of selection. Yet the calculations for multistage surveys like NSFG allow only the first stage of selection in the variance estimation. This feature of the variance estimation puzzles some survey analysts who are familiar with survey analysis software that allows incorporation of multiple stages of selection, but are unfamiliar with the theoretical properties of variance estimation from multistage samples. Some NSFG users have asked why variables identifying additional stages of selection, such as (in NSFG) sample segments within PSUs, housing units within segments, and persons within households, are not part of the variables released in the NSFG public-use data files.

A less theoretical and perhaps more accessible explanation suggests that variance estimation typically involves the calculation of a deviation between a value and a mean of the values, and then, for all intents, an averaging of the squares of the deviations. In the case of a one-stage sample, the deviation needed is between the mean of a characteristic across all elements in a cluster and the mean of the same characteristic across all elements across all clusters. That is, the variance estimation is done at an aggregate cluster level.

In that aggregate-level calculation, a sample mean is contrasted with another sample mean. The deviation between the cluster means and the overall mean is an estimate of the actual variance among clusters. However, the cluster mean is based on a sample, and because of this feature, it is subject to variability within the cluster. The sample that the cluster mean is based on has variability due to any and all subsequent stages of selection. Consequently in the NSFG, the cluster means used in the first-stage variance estimation are affected by variability among sample segments, among housing units within segments, and among persons with households. The sample cluster mean has in its very nature, variability due to second, third, and fourth stages of selection.

Hence, when the sampling variance is estimated using first-stage units, where the cluster mean is estimated from sample data, it has ‘‘built into the deviations’’ variability between first-stage or PSU-level units, between second-stage or segment-level units, and so on. It is not necessary to compute estimates separately for each stage because they are already captured almost entirely in the estimate of the first-stage variance.

Given the theoretical and this empirical explanation, it remains to show that in fact, sampling variance estimated from the first-stage units only is equivalent empirically to that obtained from a calculation that formally computes average square deviations among second- and lower-stage units. The following discussion reveals that this result does hold empirically. And as a result, the National Center for Health Statistics provides only the first-stage sampling error codes (strata and clusters) for variance estimation in the 2006–2010 NSFG, and in all previous cycles of the NSFG. This practice reduces the number of variables in the data file, eases the task of the analyst, and maintains a more secure level of data confidentiality. And the resulting variance estimates are indistinguishable from those that use additional stages of selection in the estimation process.

These arguments can be seen most clearly in the Taylor Series Linearization method of variance estimation, which is the default variance-estimation technique for complex samples in many popular statistical software packages. Many survey data sets include a series of replicate weights that enable replicated variance estimation methods like the jackknife, balanced, or bootstrapped

repeated replication. These replicate weights, however, are based on replicates that are built from the first-stage sample cluster codes. In effect then, even the repeated replication methods of variance estimation use the same principle of computing estimates of sampling variance using only codes of first-stage units.

Survey estimation software designed for analysis of complex sample survey data such as the SURVEY procedures in SAS do not allow users to specify lower-stage sampling error codes for variance estimation. Other software systems such as Stata and its svy procedures, SUDAAN, or the statistical software system R and its survey procedures, allow users to specify sample design information at multiple levels of a multistage sample design. Survey data producers seldom release public-use data files that contain the lower-stage codes needed for the multistage approach to sampling variance estimation.

What follows is an example analysis from the 2006–2010 NSFG that uses the Stata software system. The example shows how the lower-stage units in the multistage design could, if publicly available, be used to estimate sampling variances. Further, the analysis shows that the contribution of lower-stage clusters to variance estimates is at best negligible in NSFG, and does not need to be accounted for in practice by analysts.

The NSFG sample design involves selection of PSUs within strata at the first stage (see reference 2). PSUs are largely counties or groups of counties.

Within each selected PSU, there are several additional stages of selection (2). This example examines the contribution to variance estimation of only the second stage of selection, or the selection of area segments within PSUs. (Area segments were based on census blocks or groups of small census blocks.) Housing units and respondents were eventually selected within blocks as subsequent stages, but for the sake of simplicity, only the second stage or segment level contribution to sampling variance estimation will be considered.

PSUs were grouped into pairs for purposes of variance estimation. These pairings involved techniques referred to in the survey sampling realm as collapsing strata and combining strata. Although there were 110 PSUs in the sample selection, the collapsing and combining techniques resulted in 152 sampling error computing units (SECUs), or clusters, to be used in variance estimation. To protect respondent confidentiality, the SECU identification numbers were randomly ''scrambled'' to mask the identity of any given SECU.

In the typical sampling error estimation application, these SECU codes would be sufficient for appropriate variance estimation. But this example also considers the area segments selected within each first-stage SECU. These segments are lower-level clusters. Subsequently, both a single- and a two-stage variance estimation procedure are used. The single-stage approach uses the SECUs as clusters, and the two-stage approach uses the SECUs and the segment-level coding to fully account for the lower stage of selection in variance estimation. The segment-level variable is not available in the NSFG public-use data set, but was available to the authors of this report for the purposes of comparing alternative variance estimation methods.

The following variables were available for all female interviews in Cycle 7 of NSFG, a sample size of n = 12,279:

- Final sampling weight (WGTQ1Q16)
- First-stage sampling error stratum codes (SEST)
- First-stage sampling error computation unit codes (SECU)
- Second-stage segment codes within SECUs (SEGMENT)
- Ever been married (EVRMARRY)
- Ever had sex with a male (HADSEX)
- Ever used the pill, lifetime (PILL)

Calculations were computed using Stata. The Stata svyset command specifies the design variables for the subsequent svy analysis procedures. Consistent with the NSFG sample design, a negligible first-stage, within-stratum sampling fraction of 0.0001, and the corresponding finite population correction, also were used in Stata. The finite population correction is added to the calculations so that Stata recognizes that the PSUs were selected without replacement at the first and at all subsequent stages of selection specified in the svyset command. The svyset specification was as follows:

```
gen fpc1 = 0.0001
svyset secu [pweight=wgtq1q16],
strata(sest) fpc(fpc1) || segment.
```

The syntax includes the first-stage sampling error computation units (SECU) and the second-stage units (SEGMENT). The second-stage sampling units are identified after the || specification, indicating a lower level of clustering where segments are ''nested'' within levels of SECU.

To estimate proportions for the three NSFG variables of interest here (EVRMARRY, HADSEX, and PILL), the default Taylor Series Linearization method was used in the following statement:

```
svy: prop evrmarry hadsex pill.
```

The resulting standard error estimates for each level of these three variables is given in Figure I.

Next, a second svyset command was given to identify the first-stage sampling error codes only:

```
svyset secu [pweight=wgtq1q16],
strata(sest).
```

No finite population correction or second-stage sampling unit was specified. Once again, the default Taylor Series Linearization standard errors were computed for the same three estimated proportions (svy: prop evrmarry hadsex pill), and the results are shown in Figure II.

Comparing the standard errors between the two figures reveals that the estimated standard errors are virtually identical. There is minimal added contribution of the lower stage of cluster sampling to the overall variance estimates. This equivalence is not dependent on the size of the first-stage sampling fraction and finite population

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      56        Number of obs    =     12279
Number of PSUs   =     152        Population size  =  61754741
                                  Design df        =        96

------------------------------------------------------------------
             |             Linearized
             |  Proportion   Std. Err.     [95% Conf. Interval]
-------------+----------------------------------------------------
evrmarry     |
           0 |    .467186    .009175      .4489738    .4853982
           1 |    .532814    .009175      .5146018    .5510262
-------------+----------------------------------------------------
hadsex       |
           1 |   .8659312   .0069763      .8520834     .879779
           2 |   .1340688   .0069763       .120221    .1479166
-------------+----------------------------------------------------
pill         |
           1 |   .7277168   .0078266      .7121812    .7432524
           5 |   .2720054   .0078449      .2564334    .2875774
           8 |   .0002778   .0001733     -.0000662    .0006218
------------------------------------------------------------------
```

Figure I. Estimated standard errors for estimated proportions from three categorical survey variables computed using first- and second-stage components in the sampling variance estimation in the 2006–2010 National Survey of Family Growth

correction. Even when the fraction is increased from 0.0001 to 0.01 (not shown here), the results are virtually identical between the methods. Even adding a sampling fraction at the second stage (not shown here) yielded the same equivalence.

There is no guarantee that these empirical results for only three variables will hold for all variables in NSFG, or across NSFG cycles. But sampling theory suggests that they should hold across variables and across similar kinds of multistage surveys like those used in NSFG. Thus, analysts concerned about the contributions of cluster sampling at lower stages of a multistage design to variance estimates can be assured that using only the first-stage cluster identification in sampling variance estimation leads to the same result as calculations involving additional stages of selection.

```
(running proportion on estimation sample)

Survey: Proportion estimation

Number of strata =      56        Number of obs    =     12279
Number of PSUs   =     152        Population size  =  61754741
                                  Design df        =        96

--------------------------------------------------------------
             |             Linearized
             | Proportion   Std. Err.     [95% Conf. Interval]
-------------+------------------------------------------------
evrmarry     |
           0 |    .467186   .0091751      .4489737    .4853984
           1 |    .532814   .0091751      .5146016    .5510263
-------------+------------------------------------------------
hadsex       |
           1 |   .8659312   .0069764      .8520832    .8797792
           2 |   .1340688   .0069764      .1202208    .1479168
-------------+------------------------------------------------
pill         |
           1 |   .7277168   .0078267       .712181    .7432526
           5 |   .2720054    .007845      .2564332    .2875776
           8 |   .0002778   .0001733     -.0000662    .0006218
--------------------------------------------------------------
```

Figure II. Estimated standard errors for estimated proportions from three categorical survey variables computed using the first-stage component in the sampling variance estimation in the 2006–2010 National Survey of Family Growth

Vital and Health Statistics Series Descriptions

ACTIVE SERIES

Series 1. **Programs and Collection Procedures**—This type of report describes the data collection programs of the National Center for Health Statistics. Series 1 includes 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**—This type of report concerns statistical methods and includes analytical techniques, objective evaluations of reliability of collected data, and contributions to statistical theory. Also included are experimental tests of new survey methods, comparisons of U.S. methodologies with those of other countries, and as of 2009, studies of cognition and survey measurement, and final reports of major committees concerning vital and health statistics measurement and methods.

Series 3. **Analytical and Epidemiological Studies**—This type of report presents analytical or interpretive studies based on vital and health statistics. As of 2009, Series 3 also includes studies based on surveys that are not part of continuing data systems of the National Center for Health Statistics and international vital and health statistics reports.

Series 10. **Data From the National Health Interview Survey**—This type of report contains 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. Series 10 is based on data collected in this 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**—In this type of report, 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 13. **Data From the National Health Care Survey**—This type of report contains 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 20. **Data on Mortality**—This type of report contains 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**—This type of report contains 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 23. **Data From the National Survey of Family Growth**—These reports contain statistics on factors that affect birth rates, including contraception and infertility; factors affecting the formation and dissolution of families, including cohabitation, marriage, divorce, and remarriage; and behavior related to the risk of HIV and other sexually transmitted diseases. These statistics are based on national surveys of women and men of childbearing age.

DISCONTINUED SERIES

Series 4. **Documents and Committee Reports**—These are final reports of major committees concerned with vital and health statistics and documents. The last Series 4 report was published in 2002. As of 2009, this type of report is included in Series 2 or another appropriate series, depending on the report topic.

Series 5. **International Vital and Health Statistics Reports**—This type of report compares U.S. vital and health statistics with those of other countries or presents other international data of relevance to the health statistics system of the United States. The last Series 5 report was published in 2003. As of 2009, this type of report is included in Series 3 or another series, depending on the report topic.

Series 6. **Cognition and Survey Measurement**—This type of report uses methods of cognitive science to design, evaluate, and test survey instruments. The last Series 6 report was published in 1999. As of 2009, this type of report is included in Series 2.

Series 12. **Data From the Institutionalized Population Surveys**—The last Series 12 report was published in 1974. Reports from these surveys are included in Series 13.

Series 14. **Data on Health Resources: Manpower and Facilities**—The last Series 14 report was published in 1989. Reports on health resources are included in Series 13.

Series 15. **Data From Special Surveys**—This type of report contains 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. The last Series 15 report was published in 2002. As of 2009, reports based on these surveys are included in Series 3.

Series 16. **Compilations of Advance Data From Vital and Health Statistics**—The last Series 16 report was published in 1996. All reports are available online, and so compilations of Advance Data reports are no longer needed.

Series 22. **Data From the National Mortality and Natality Surveys**—The last Series 22 report was published in 1973. Reports from these sample surveys, based on vital records, are published in Series 20 or 21.

Series 24. **Compilations of Data on Natality, Mortality, Marriage, and Divorce**—The last Series 24 report was published in 1996. All reports are available online, and so compilations of reports are no longer needed.

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 5419
Hyattsville, MD 20782

Tel: 1–800–CDC–INFO (1–800–232–4636)
TTY: 1–888–232–6348
Internet: http://www.cdc.gov/nchs
Online request form: http://www.cdc.gov/cdc-info/requestform.html

For e-mail updates on NCHS publication releases, subscribe online at: http://www.cdc.gov/nchs/govdelivery.htm.

U.S. DEPARTMENT OF HEALTH & HUMAN SERVICES

Centers for Disease Control and Prevention
National Center for Health Statistics
3311 Toledo Road, Room 5419
Hyattsville, MD 20782

DHHS Publication No. 2013–1358, Series 2, No. 158
CS238921