

APPENDIX 2:
EXCERPTS FROM THE
NSFG SERIES 2 REPORT,
FORTHCOMING IN ABOUT JANUARY 2010.

Vital and
Health Statistics

**Continuous National Survey of Family
Growth: New Concepts for Sample Design
and Analysis**

by James M. Lepkowski, PhD, Institute for Social Research, University of Michigan;
William D. Mosher, Ph.D, and Karen E. Davis, M.A., NCHS; Robert M. Groves, PhD,
and John Van Hoewyk, PhD, Institute for Social Research, University of Michigan

Series 2: No. XX

Draft 12B--August 19, 2009.

*Excerpt for the NCHS Board of Scientific Counselors.
THIS REPORT IS STILL IN REVIEW*

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 women and men 15-44 years of age in the United States. This report describes the sample design for the NSFG's new continuous design, and the effects of that design on weighting and variance estimation procedures. A working knowledge of this information is important for researchers who wish to produce accurate, valid national estimates from the data.

Because the continuous NSFG was a significant departure from the previous periodic (once every 6-7 years) survey, this initial report describes the overall concepts and procedures for the sample design, weighting, and variance estimation. This report is being released before the survey results so that researchers will have the background information necessary to understand the file when the data files are released. After each data file is released from the new design, a new report in this Series will include specific results of the weighting, imputation, and variance estimation, and specific guidance on analysis techniques with each new data file. In this way, researchers will have more of the information they need in a more timely manner.

Methods — The NSFG Continuous design is based on an independent, national probability sample of women and men 15-44 years of age. Fieldwork was carried out by the University of Michigan's Institute for Social Research (ISR) under a contract with NCHS. In-person, face to face interviews have been conducted by professional female interviewers using laptop computers.

Results — Analysis of NSFG data will require 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 will be provided on the data files. The rate of missing data in the survey is generally low. However, missing data are being imputed for about 600 key variables that are used for most analyses of the survey, called "recodes." Imputation was accomplished using a multiple regression procedure with software called IVEware, available from the University of Michigan website.

Keywords: survey methodology; imputation; variance estimation; National Survey of Family Growth.

Continuous National Survey of Family Growth: New Concepts for Sample Design and Analysis

Executive Summary

The National Survey of Family Growth (NSFG) obtains detailed information on factors affecting childbearing, marriage, and parenthood from a national probability sample of women and men 15 to 44 years of age. This Series 2 report describes the procedures used to select the sample, develop the sampling weights, impute missing data, and estimate sampling errors. The report should be useful to data analysts who intend to use the NSFG public use data files, and need to understand how the sample design, weighting, imputation, and variance estimation may shape their intended analyses. It may also be useful to survey methodologists, who may wish to compare their procedures to those used in the NSFG.

This report is a significant departure from previous NSFG Series 2 reports in two ways. First, previous NSFG Series 2 reports have been released 3-4 years after data collection ended, which meant that many data analysts had to do their initial analyses before the Series 2 report was available. This report, in contrast, is being released with the data file, so that NSFG data users will have its information when they are ready to do their initial analyses. Second, in continuous interviewing, specific results for each public use data file will be released as soon as they are available, on the NSFG web site and in subsequent Series 2 reports. This procedure should deliver more information into the hands of data users in a more timely way than in past NSFG cycles.

The National Survey of Family Growth is designed and administered by the National Center for Health Statistics (NCHS), an agency of the US 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;
- the medical care associated with contraception, infertility, and childbirth;
- factors affecting marriage, divorce, cohabitation, and adoption;
- what women and men do to raise their children; and
- men’s and women’s attitudes about sex, childbearing, and marriage.

The continuous NSFG is being conducted by the University of Michigan’s Institute for Social Research under a contract with NCHS. As continuous interviewing began, a national probability sample of approximately 5,000 women and men 15-44 years of age living in households in the United States were being interviewed each year.

Continuous interviewing began about July 1, 2006. The interviews were conducted in person by trained female interviewers using laptop, or notebook computers, a procedure called computer-assisted personal interviewing (CAPI). The interviews for women averaged approximately 80 minutes and the interviews for men averaged about 60 minutes.

The continuous NSFG was based on a sampling plan that was intended to provide larger and more reliable samples than ever before in the NSFG's history. The continuous national sample is drawn from 110 major areas, or Primary Sampling Units (PSUs), divided into four national sub-samples. Each of the four sub-samples has been worked for one year, so the entire 110-PSU design is expected to be completed in a four-year period. The entire 4-year data file (mid-2006 to 2010) should yield at least 20,000 interviews in 110 PSUs for the full 2006-2010 time period, yielding the largest sample in the NSFG's history.

Black, Hispanic, teenage, and female respondents were sampled at higher rates than others. Sampling weights were used to compensate for the different sampling rates of these various groups, and for different non-response rates. Sampling errors are estimated using software that takes the weights and the stratified cluster sample design features into account. Such software is now widely available in such packages as SAS, Stata, SPSS, SUDAAN, and others.

In addition, for key variables, referred to in this and other NSFG reports as "recodes," item missing values have been replaced in the data file by predicted or imputed values. The imputed values in the approximately 600 recode variables were identified in a companion variable (or "imputation flag") that indicates whether the value for a particular case was imputed or reported.

The rest of this report describes how the sample was designed and selected, how sampling weights are computed and adjusted to compensate for the different sampling rates and other factors, and how missing data are imputed for selected recodes. This report concludes with a section on "NSFG Continuous Design: A Guide for Analysts," which includes some recommendations for those wishing to do research with the NSFG. To reiterate this advice, the section includes ten Frequently Asked Questions (or FAQ's) to clarify these practical issues for analysts further.

In general, continuous interviewing allows researchers to choose one of three approaches:

- A) to use the whole data set as one has always used the NSFG;
- B) to use a subset of the data to look at a more restricted time period; and
- C) to combine several consecutive years to study smaller subgroups and rare events.

For the most part, approaches B and C will not be available until the second

data release (including at least 4 years of data). For the first data release, most NSFG data users will continue to use the NSFG as a cross-section, while planning future analyses that will allow them to use the full time series capabilities of the data set.

This report is designed to describe the continuous NSFG sample design at three levels of detail:

- first, this Executive Summary is for readers seeking a general, non-technical understanding of the survey procedures;
 - second, summaries for each major section of the text are included to provide somewhat more information on each design feature;
 - and third, in the rest of the report, the full technical details are provided for the interested reader.
-

Design Specifications

The NSFG's continuous sample design was based on the following objectives:

- The target population for the continuous NSFG was to be the household population of men and women aged 15-44 in households in the 50 states and the District of Columbia.
- Screening interviews were to be conducted in each sampled household, to determine if anyone 15-44 years of age lived there, and if so, to select one person from the household for the NSFG interview.
- The selection was to be random but to sample certain sub-groups at higher rates. Those groups included teenagers (15-19 years of age), Hispanic men and women, and non-Hispanic black men and women.
- Data collection was to be conducted only by in-person, face-to-face interviewing, with the respondent's privacy and confidentiality ensured.
- 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.
- Given the sensitivity of the interview content, all interviewers were to be female. That is, men and women were to be interviewed by a female interviewer.

- Data collection was to be completed using Computer-Assisted Personal Interviewing (CAPI). One section of the questionnaire was to be administered using Audio Computer-Assisted Survey Interviewing (ACASI), in which the respondent would listen to a pre-recorded audio reading of the questions with a headset connected to the laptop. The questions were also to be displayed on the computer screen. Survey instruments were to be programmed on laptop computers meeting specified requirements.
- The available funds appeared to allow a sample of approximately 4,400-5,000 interviews per year, yielding 17,000 – 20,000 interviews over a 4-year data collection period. Of these, about 45% were to be males and 55% females; about 20% teenagers 15-19; about 20% Hispanic; and about 20% black or African American.
- The following sample sizes and design effects were expected over a 4-year period:

	<u>Sample N</u>	<u>Design Effect</u>
Total	7,600 – 20,000	3.9
15-19	3,500 – 4,000	1.5
Hispanic	3,500 – 4,000	1.6
Black/African American	3,500 – 4,000	1.6
Male	7,900 - 9,000	2.2
Female	9,700 - 11,000	2.8

- The sample was to be designed so that any number of individual years could be cumulated to form a nationally representative sample for analysis. Practical considerations, such as sample size and design effects, however, made it likely that at least two years of data would be needed for most analytical purposes.
- Signed informed consent was required for every selected eligible respondent 15-44 years of age. Minors 15-17 years of age were required to have the signed consent of a parent before being asked for their own signed assent.
- Interviewers were to ask questions about fertility, contraceptive use, sources and types of family planning services, and maternal and child health using structured questions similar to those used in the 2002 (Cycle 6) NSFG, and read by an interviewer (in the CAPI part of the interview) or heard on a headset (in the Audio CASI part of the interview).
- The contractor, in cooperation with NCHS, was to design and implement procedures for measuring and controlling the quality of data collection and data preparation procedures, including verification of a sample of interviews.

NSFG Continuous Design: A Guide for Analysts

NSFG continuous design allows the release of the data more frequently than in the past, as it is collected. At the time of this writing, continuous NSFG data are expected to be released for the period beginning approximately July 1, 2006, through December 2008, a 30 month data collection period, and for the period beginning approximately July 1, 2006, through June 2010, a 48 month period.

This kind of more frequent release means NSFG users have analytic opportunities which were not possible with previous releases of NSFG data in cycles. These opportunities are the consequence of collecting data over an extended time period in smaller national samples in several consecutive years, as described elsewhere in this report. The smaller national samples can be accumulated to form larger national samples over longer periods of time. Those larger national samples can be analyzed in at least three ways that are of interest to NSFG users.

First, the entire continuous NSFG data file can be analyzed as a whole, just as previous NSFG data files were analyzed. Second, the NSFG releases will allow users to compute estimates for groups of calendar years. And third, longer time periods allow users to obtain larger samples of smaller groups in the population, or less frequent behaviors reported by respondents, to increase sample size to do more reliable analyses that were not possible with previous cycles of NSFG.

Analyzing the Continuous NSFG as Periods of Data Collection. The present plan is that under the sample design described in this report, there would be two releases of NSFG continuous design data. The first would include over 13,000 interviews collected during the 30-month period beginning approximately July 1, 2006, through December 31, 2008 (“the 2006-2008 data”), and the second will include over 20,000 interviews collected during the 48-month period beginning approximately July 1, 2006, through June 30, 2010. Thus, users will have two sets of data of increasing sample size to work with during the course of the current continuous design.

For each of these releases of data, users can treat the data much as they might have handled previous NSFG cycles. That is, a user could analyze the 2006-2008 data as the next cycle of the NSFG. The structure of the continuous NSFG data is very similar to Cycle 6 NSFG, and there is a weight and sampling error codes that allow the user to compute estimates for the entire 30 month period during which data were collected. Users can compute estimates and fit models as they did for Cycle 6, or earlier cycles.

But there is one important difference, the reference time period. For Cycle 6, for example, data were collected over an 11 month data collection period during 2002 and through the beginning of 2003. One weight and a set of sampling error codes were provided, and users referred to the findings as being about the US population in 2002.

With the first release of the NSFG continuous design, a user could repeat an analysis done on Cycle 6, with a single weight and set of sampling error codes. However, the user will have to refer to, the US population over the *period mid- 2006 through 2008*. That is, the time period of inference changes under the NSFG continuous design. The 2006-2008 data set weight is 'centered' in the middle of this time period. The population for which inferences can be made will thus be the *average* US population over the period from approximately July 1, 2006, through December, 31, 2008, for the first anticipated release, or from beginning approximately July 1, 2006, through June 30, 2010, for the second. Again, a single weight and set of sampling error codes will be available for the entire four year period in the latter file. Inferences will be about the *average* population over the two and one-half year or the four year period. The usual reference period implicit in Cycles 1-6 analyses changes under the NSFG continuous design from one year to more than one year.

The continuous NSFG user gets larger sample sizes than ever before that were obtained at lower cost per completed interview by a continuous design, but the user must refer to periods of time longer than one year. The larger sample sizes make more detailed study of the entire sample or subgroups possible than before, where subgroup analysis is done just as in previous NSFG cycles. To the analyst, there is no change in how they go about analysis, and they have more frequent releases of data, with successively larger sample sizes, and a single weight and set of sampling error codes for each release.

Analyzing the Continuous NSFG for Shorter Periods of Time. The continuous design has another advantage: as more continuous NSFG data become available, users will also be able to compute estimates for shorter time periods within each data file released.

For example, in a future release, it may be possible to make estimates for periods of two consecutive calendar years (2007-8 and 2009-10) so that NSFG estimates may be (a) compared to other data sources (such as vital statistics and other surveys), or (b) used for a more familiar time period. However, for a two-year time period, the single weight for the entire 30-month or 48-month release will no longer be sufficient. For a sample of 24 months of interviews, sample sizes will be smaller than for 30 or 48 months, and different weights will be required. When at least two 24-month time periods are available for analysis, weights will be provided so that their results can be compared.

Questions added in 2007--In addition, new questions were added to the NSFG continuous design questionnaire in July, 2007. The variables generated by these new questions are contained in the first anticipated 2006-2008 release, and were collected for an 18 month period from July, 2007, through December, 2008. Yet, any analysis conducted which uses these 2007 variables will have missing or 'inapplicable' codes for respondents interviewed before the questions were added to the survey. Sample sizes will be reduced accordingly, from approximately 13,500 to about 8,000, and weighted

numbers using the 2006-2008 weight will be too small. To compensate, a separate weight will be needed for the variables introduced in 2007 (“2007 variables”) so that NSFG data users can analyze them. Of course, the expected 2006-2010 data will allow the “2007 variables” to be used for a period of 36 months (containing an expected 15,000 interviews).

NSFG staff are examining the properties of estimates from the 2007 variables in the 2006-2008 release as of this writing. It is expected that a weight for those variables (based on 18 months of interviews) will be included on the file, but greater caution will be required because those statistics will be based on about 8,000 interviews (about 4,400 women and 3,600 men) instead of the 13,500 in the full dataset.

There are of course several shorter time periods within the expected NSFG releases that could be selected by a user. However, as the time period becomes shorter, such as moving to a single year or even a calendar quarter within a year (such as the period January through March, 2007), the sample sizes become smaller, and estimates become less reliable. At a certain level, the NSFG estimates will be based on such small sample sizes that any estimates produced will no longer have adequate precision.

As a result, NSFG does not recommend that estimates be computed for intervals shorter than 18 months. Calendar quarter estimates should not be computed, and weights to allow users to compute such estimates will not be provided.

At the present time, NSFG staff are examining whether estimates for single calendar years will be adequately reliable for any purpose. A calendar year contains approximately 5,000 completed female and male interviews. Given the NSFG’s oversampling of minority groups, teenagers, and women, and its clustered design, this sample size is at the borderline of acceptable levels of precision. For the anticipated first release of 2006-2008 data, which contains only the two calendar years 2007 and 2008, calendar year estimates are not recommended. No calendar year weights will be provided in the first data file release.

For the second expected release of 2006-2010 data, three calendar years could be covered (2007, 2008, and 2009). Between the first and second anticipated releases, NSFG staff will be carefully examining the precision of estimates for calendar year periods. If the precision is deemed to be acceptable by NCHS standards, calendar year weights for 2007, 2008, and 2009 will be provided in the 2006-2010 release data. Again, the same set of sampling error codes that will be used for the period 2006-2008, or 2006-2010, will be used for these shorter single calendar year periods.

Once several calendar-year time periods of NSFG data become available, users may then compare estimates across time. Guidance on how to make comparisons of these time periods will be provided in the user guide for the 2006-2010 release. If single calendar year weights are provided in the second release, users will be able to

compare estimates for 2007, 2008, and 2009. Again, precision for comparisons will be limited, due to the smaller sample sizes, and differences will have to be quite large to be statistically significant.

Analyzing Continuous NSFG for Small Subgroups or Rare Behaviors. The continuous NSFG allows users the opportunity to accumulate data across several years of data collection in order to study small subgroups or rare behaviors. Accumulation is useful when a short time period simply does not provide enough cases for adequate levels of precision in estimates. In the expected 2006-2010 release, users will (with the overall larger sample size) be able to select proportionately smaller subgroups for analysis than possible with previous NSFG cycles.

Yet, there will be limits to the size of sample that should be analyzed. Continuous NSFG release weights and the single set of sampling error codes provided will allow users to make time period specific inferences about smaller samples selected from the full release of data. The user must be cautious, and exercise professional judgment, about when the levels of precision for a given subgroup are simply not adequate, just as they have had to do with previous cycles of NSFG (or any other survey). Guidelines regarding the minimally acceptable levels of precision for such subgroup analyses have not yet been determined at this time. However, NSFG staff themselves exercise caution in publishing or releasing findings for estimates of prevalence or averages whenever a standard error is more than 25% the size of the estimate. Estimated proportions or means where the standard error is larger than 25% of the estimate itself, or where the numerator or denominator are too small, must be interpreted with caution.

Additional Time Periods. Besides single calendar years, other time intervals may be considered. However, in order to provide comparability with other survey data, which are most often released for calendar year estimates, the NSFG staff only expects to provide weights for calendar year periods (two-year and possibly single year in the second release). There are no plans at the time of this writing to provide weights for intervals covering data collection years, such as from July, 2006, through June, 2007.

Table C. Expected Design Features of Continuous NSFG, 2006-2010

Years of Data Collection	No of PSU's	Months Of Data Collection	Weights Provided on file?	No. of SECU's	DF for Variance Estimates
FIRST RELEASE, 2006-2008					
2 ½ year	85	30	Yes	122	71
2-year	85	24	No	NA	NA
1 ½ year	58	18	Yes ^a	122	71
1 year	33 ^b	12	No ^c	92	46
¼ year	33 ^b	3	No	NA ^c	NA ^c
SECOND RELEASE, 2006-2010					
4 year	110	48	Yes	152	96
2-year	85	24	Yes	122	71
3 year	85	36	Yes	152	96
1 year	33 ^b	12	Yes ^a	92	46
¼ year	33 ^b	3	No	NA ^c	NA ^c

^a Weight will be provided if research shows that precision is adequate.

^b 35 PSU's in first data collection year.

^c Not Applicable – estimates unreliable.

PSU's = Primary Sampling Units

SECU's = Sampling error computing units

DF = Degrees of Freedom

FAQ's: Summary in Question-Answer Format

To clarify the current outlook for analysis using the NSFG, the following section shows some Frequently Asked Questions (FAQ's) and their answers. Additional information will be released on the NSFG web site (<http://www.cdc.gov/nchs/nsfg.htm>) as it becomes available.

1. Given that the continuous interviewing design uses quarterly samples, can I analyze the data for just one quarter?

No. Sample sizes for a single quarter are too small to provide estimates with adequate levels of precision. In general, NSFG staff recommend using data for at least 18 months for adequate precision. In the first data file, weights will be included only for the full 30 months of interviewing, and for the last 18 months of interviewing (when certain new questions were introduced). **Analysts should only use time periods for which sampling weights are provided.**

2.The public use data file has both recoded variables (“recodes”) and raw variables. Which ones should I use?

NCHS recommends using recodes, when they are appropriate for a given analysis, for two main reasons. First, the recoded variables have been studied for consistency and any missing data has been resolved through imputation. Thus, using the recodes allows the analyst to make use of the intensive scrutiny given to them. Second, NCHS uses the recoded variables used in many of the tables in NCHS official publications, so the user can check their own results with those of NCHS to ensure comparability. If a recoded variable is not available for a specific analysis, then the raw variables or computed variables should be used.

3.Will the continuous interviewing design use different questionnaires over the years?

Changes occurred at the beginning of the 2nd and 3rd years of interviewing (no changes were made for year 4). But over 95% of the questions remain the same from year one to year four of continuous interviewing. Variables measured in only some of the years will be noted in the public use documentation of the data sets.

4.How do I combine the different quarters of data collection in my analysis?

The NSFG staff and contractor will include sampling weights on the data files for specific years, or groups of quarters, for which analysis is appropriate. See the data release documentation for full guidance on your specific analysis.

5.Given that the size of the data set can become large over the many quarters of continuous interviewing, can I analyze the data separately for different states in the country?

Although the number of data records in the pooled data set of the continuous survey can become large when several years of data are combined, the sample is limited to 110 primary areas. These primary areas do not fall in all states. As long as this design is used, estimates cannot be computed for individual states. If the sample design were modified and sample sizes increased, estimates for at least some states would become possible.

6. Can I combine the data for males and females in continuous interviewing?

Yes. There is a combined file for males and females, which contains variables that are common to the two sexes. Using the sampling weights, estimates for both males and females can be made.

7. For a recoded variable, how do I find out what questions in the questionnaire contributed to it?

The “Recode specifications” are given in the public use data file documentation. These specifications show how each recoded variable was constructed. The NSFG staff recommends using recodes when they are appropriate for a particular analysis, because they have been checked for accuracy, edited, and imputed.

8. Should I obtain the same results on birth-related statistics from analyzing continuous NSFG as vital statistics?

Birth statistics based on the National Vital Statistics System are derived from a complete count of the approximately 4 million birth certificates filed each year. The continuous NSFG, in contrast, estimates these births with a sample of a few hundred births each year to women in the continuous NSFG sample. Therefore, continuous NSFG estimates will not match those from the birth certificates exactly, primarily because of sampling error. But there are also some differences between the coverage, nonresponse, and measurement features of the two approaches to estimation.

9. What format is used for the public use data set?

The data sets are released in ASCII format and are compatible with several major statistical software systems that permit analysis of complex survey data, including SAS, SPSS, and Stata.

10. How is the continuous pooled data set any different from analyzing any other sample survey?

For the most part, analyzing the pooled data is no different than analyzing a previous NSFG cycle, at least in terms of the application of standard survey estimation software to the data file. Each continuous NSFG data file can be thought of and treated as a single cycle of the NSFG, but the time period to which the results apply must be described accurately.

The continuous NSFG allows a user to obtain 'period prevalence' estimates for the population over a multi-year period. Previous NSFG cycles allowed the user to report estimates for a single year. The continuous NSFG will allow the user to report estimates for a period of several years. For example, the proportion of women 15-44 who are currently using the oral contraceptive pill may be reported, once the 2009-2010 data are released, for the period from 2006 to 2008, or the period 2009 to 2010, or for the period 2006 through 2010. The first two of these periods may be reported by a user interested in contrasting rates across the US population in each of those time intervals. The third period allows the user to report the rate across a longer time interval, but with greater precision, since the sample size of the pooled data will be larger.

Thus, the continuous NSFG requires reporting results for longer time periods, or for contrasting time periods. It may also be cumulated over time to produce estimates with greater precision. For example, proportion of Hispanic females using one type of contraception for the period 2006 to 2008 will have about one-half the sample size of the same proportion for the period 2006 to 2010. Cumulating over a longer time period will provide more precise estimates, which will be especially useful for smaller subgroups of interest, at the price of a longer time period reference for the estimate.

In all these estimates, the analyst must use appropriate weights and variance estimates for the specific time period used. The appropriate weights and sampling error codes will be provided with each release of the continuous NSFG that allow proper estimation for different time periods of interest. The presence of alternative time period weights will require users to choose the time period of estimation and the appropriate weight to apply, but the benefits include greater analytical flexibility and larger sample size.