2013-2015 NSFG USER'S GUIDE APPENDIX 2:
SAS AND STATA SYNTAX GUIDELINES FOR COMMON FILE MANIPULATIONS

In this appendix are SAS and Stata syntax guidelines for three main types of file manipulations that are commonly done with the NSFG public use data. Along with these syntax guidelines, some further technical guidance is included on conducting statistical analyses with NSFG data combined across survey years.

These guidelines and examples are specified assuming that you are using SAS and Stata datasets, based on the program statements provided on the NSFG webpage. If you are working directly with the raw (ASCII) data files, you will need to adapt these programs to use the "INFILE" and "INPUT" statements in SAS or the "INFIX" statement in Stata. Also, the examples are provided in a generic format. You must adapt to your own local computing environment with regard to file names, file paths, and libnames, if applicable.

For further guidance on variance estimation using NSFG data, see also the section on **SAMPLE WEIGHTS AND VARIANCE ESTIMATION** in Part 1 of the User's Guide. Examples of syntax and output for three selected variance estimation scenarios are posted on the NSFG webpage.

Combining Data from Female Respondent and Female Pregnancy Files within 2013-2015 NSFG:

Note:
The information provided below is essentially the same as provided in the 2011-2013 NSFG User's Guide Appendix 2. The syntax examples and selected details have simply been updated to apply to the 2013-2015 NSFG.

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

To facilitate analysis based on women as the units of analysis, selected pregnancy-specific variables were placed on the Female Respondent file for each of up to 20 pregnancies. (No female respondent in the 2013-2015 NSFG reported more than 20 pregnancies.) See File Indexes in **Appendix 1 of the User's Guide** for basic file layout information and the program statements posted in the 2013-2015 NSFG webpage for precise column locations on the ASCII data files.

Pregnancy-specific variables in the female respondent file include the following recode variables and imputation flags:

- pregnancy outcome (OUTCOM01-20)
- date pregnancy ended (DATEND01-20)
- age of woman at time of pregnancy outcome (AGEPRG01-20)
- formal marital status at pregnancy outcome (MAROUT01-20)
- informal marital status at pregnancy outcome (RMAROUT01-20)
- date of conception (DATCON01-20)
- age of women at time of conception (AGECON01-20)
- formal marital status at conception (MARCON01-20)
- informal marital status at conception (RMARCON01-20)
- current living situation of 1st liveborn child from the pregnancy (LIVCHILD01-20)
- wantedness of pregnancy by R at time of conception (Cycle 4 definition) (OLDWR01-20)
- wantedness of pregnancy by H/P at time of conception (Cycle 4 definition) (OLDWP01-20)
- wantedness of pregnancy by R at time of conception (Cycle 6 definition) (WANTRP01-20)
- wantedness of pregnancy by H/P at time of conception (Cycle 6 definition) (WANTP01-20)
- detailed wantedness of pregnancy by R at time of conception (NWWANTRP01-20)

In addition, to facilitate analyses based on pregnancies as the units of analysis, some key respondent-specific characteristics were included on the pregnancy (interval) file. These variables have "respondent-level variable" or "respondent-level recode" affixed to the end of their universe statements in the codebook. The variables include:

Questionnaire data

- Century-month of R's birth (cmbirth)
- Age at time of household screener (agescrn)
- current pregnancy status and gestational length of a current pregnancy (HOWPREG_N, HOWPREG_P, moscurrp, NOWPRGDK)
- nativity status (whether born outside U.S.) and year when she came to the U.S. to stay (BRNOUT, YRSTRUS)

Recodes

- age at interview (AGER)
- formal marital status at interview (FMARITAL)
- informal marital status at interview (RMARITAL)
- race, regardless of Hispanic origin (RACE)
- Hispanic origin (HISPANIC)
- race and Hispanic origin using 1977 OMB standards (HISPRACE)
- race and Hispanic origin using 1997 OMB standards, for multiple race reporting (HISPRACE2)
- whether R is currently pregnant (at interview) (RCURPREG)
- number of pregnancies ever had (by time of interview) (PREGNUM)
- number of liveborn children (by time of interview) (PARITY)
- religious affiliation at interview (RELIGION)
- education at interview (EDUCAT, HIEDUC)
- health insurance coverage status at interview (CURR_INS)
- poverty level of household's income at interview (POVERTY)
- receipt of public assistance in the last year (PUBASSIS)
- labor force status at interview (LABORFOR)
- metropolitan residence at interview (METRO)

Analyses using the pregnancy (interval) file may require additional information about women from the respondent file, and analyses using the respondent file may require additional information about pregnancies from the pregnancy (interval) file. Using the common case identification number (CASEID), and the pregnancy number (PREGORDR), the pregnancy (interval) and respondent files can be merged to produce a file containing both respondent information and pregnancy information. The resulting file can be either respondent-based (up to 5,699 records) or pregnancy (interval)-based (up to 9,358 records). See examples below for examples of SAS and Stata code that will allow you to merge the respondent and pregnancy files either way.

One additional note about sample design and weight variables:
These variables have the same names on both the female respondent and female pregnancy files, and should require no renaming when you combine data from these files. See section on **SAMPLE WEIGHTS AND VARIANCE ESTIMATION** in Part 1 of User's Guide for further details.

- SEST
- SECU
- WGT2013_2015

Adding Respondent Variables to a Pregnancy (Interval) Based File

The SAS and Stata template programs below will yield a pregnancy-based SAS or Stata data file with a maximum of 9,358 records if no subsetting of pregnancy records is done. The respondent-based variables that are not already on the pregnancy file will be added to EACH pregnancy record with the same CASEID (case identification number). If you wish to subset pregnancy records as part of this merge, you would use a line such as the one shown in red in each program template.

In addition to CASEID, the following template programs refer to OUTCOME, a pregnancy file recode indicating the outcome of each pregnancy reported by female respondents. It has a value of 1 for live birth, values 2-5 for different categories of non-live birth outcomes, and a value of 6 for a current pregnancy at time of interview. *(See Appendix 3b or the Webdoc codebook documentation for detailed specifications for the OUTCOME recode.)*

Using SAS:

```
/*Select variables from the female respondent file*/
      DATA RESP; SET RESPONDFILE
            (KEEP= CASEID [other variables you wish to include]);
      RUN;

/*Select variables from the pregnancy file*/
      DATA PREG; SET PREGFILE
            (KEEP= CASEID OUTCOME [other variables you wish to include]);
      IF OUTCOME=1; /*subsetting only those pregnancies ending in live birth */
      RUN;

/*Sort both RESP & PREG sasfiles by the merge variable, CASEID */
      PROC SORT DATA=RESP; BY CASEID; RUN;
      PROC SORT DATA=PREG; BY CASEID; RUN;

/* Merge the 2 sorted files, using the pregnancy file as the driver of the merge.
This is accomplished using the "in=a" following "PREG" */
      DATA ALLPREG; MERGE RESP PREG (IN=A);
            BY CASEID;
            IF A;
```

```
RUN;

/* The above merge will produce a 2013-2015 sasfile with:
      9,358 records if the line in red is NOT used, and
      6,489 records if the line in red is used to subset only live births */
```

Using Stata:

```
* Select variables from the female respondent file and sort by the merge variable,
CASEID
      use RESPOND
      keep CASEID [other variables you wish to include]
      sort CASEID
      save RESPONDSORTED, replace
      clear

* Select variables from the pregnancy file and sort by the merge variable, CASEID
      use PREG
      keep CASEID OUTCOME [other variables you wish to include]
      keep if OUTCOME==1 * this line subsets only those pregnancies ending in live
                           birth
      sort CASEID
      save PREGSORTED, replace

* Merge the 2 sorted files
      merge m:1 CASEID using RESPONDSORTED
      keep if _merge==3
      save RESPPREG, replace

* The above merge will produce a 2013-2015.dta file with:
*     9,358 records if the line in red is NOT used, and
*     6,489 records if the line in red is used to subset only live births
```

Adding Pregnancy Variables to a Respondent Based File

The SAS and Stata template programs below will yield a respondent-based SAS or Stata data file with selected pregnancy file variables merged in. Though the respondent file includes information for 5,699 women, not all of them have ever been pregnant, so the maximum records you could have based on ever-pregnant women is 3,476 (see recode PREGNUM in **Appendix 3a, Female Respondent Recode Specifications**). The examples below show how to subset respondents who have had at least 1 live birth, and output a dataset with only the pregnancy file information for their most recent live birth. Such a subsetted file may be helpful if you wish to examine, for example, breastfeeding for the most recent birth in the context of the respondent's contraceptive, work, or relationship experiences.

In addition to CASEID and OUTCOME described above, these template programs reference the PREGORDR variable, which indicates the pregnancy order or number (that is, 1st pregnancy, 2nd pregnancy, etc.).

Using SAS:

```
/*Select variables from the female respondent file*/
      DATA RESP;
      SET RESPONDFILE (KEEP= CASEID [other variables you wish to include]);
      RUN;

/*Select variables from the pregnancy file*/
      DATA PREG;
```

```
      SET PREGFILE (KEEP= CASEID PREGORDR OUTCOME
                  [other variables you wish to include]);
      IF OUTCOME=1; /*subsetting only those pregnancies ending in live birth*/
      RUN;

/*Sort PREG sasfile by CASEID*/
      PROC SORT DATA = PREG;
      BY CASEID;
      RUN;

/* Keep only the last live birth for each respondent (CASEID) */
      DATA LASTPREG;
            SET PREG; BY CASEID;
            IF LAST.CASEID THEN OUTPUT;   /* only 1 record output per CASEID */
            RUN;

/*Sort both RESP & PREG sasfiles by the merge variable, CASEID */
      PROC SORT DATA=RESP;
            BY CASEID;
            RUN;
      PROC SORT DATA=LASTPREG;
            BY CASEID;
            RUN;

/* Merge the 2 sorted files, using LASTPREG as the driver of the merge */
/* This is accomplished using the "in=a" following "LASTPREG" */
      DATA LASTBRTH;
            MERGE RESP LASTPREG (IN=A);
            BY CASEID;
            IF A;
            RUN;

/* The above merge will produce a 2013-2015 SAS file with 3,067 records, which is
consistent with the respondent-based recode, LBPREGS, showing how many respondents
had at least 1 pregnancy resulting in live birth (see Appendix 3a). */
```

Using Stata:

```
* Select variables from the female pregnancy file and sort by the merge variable,
CASEID
      use PREG
      keep CASEID PREGORDR OUTCOME [other variables you wish to include]

* This line subsets only those pregnancies ending in live birth
      keep if OUTCOME==1
      gen LAST=1 if CASEID!=CASEID[_n+1]
      keep if LAST==1
      sort CASEID
      save LASTPREG, replace
      clear

* Select variables from the female respondent file and sort by the merge variable,
CASEID
      use RESP
      keep CASEID [other variables you wish to include]
      sort CASEID
      save RESPSORTED, replace

* Merge the 2 sorted files
      merge 1:m CASEID using LASTPREG
```

```
keep if _merge==3
save LASTBIRTH, replace

* The above merge will produce a 2013-2015 SAS file with 3,067 records, which is
consistent with the respondent-based recode, LBPREGS, showing how many respondents
had at least 1 pregnancy resulting in live birth (see Appendix 3a).
```

Combining Data for Males and Females within 2013-2015 NSFG

Note:
The information provided below is essentially the same as provided in the 2011-2013 NSFG User's Guide Appendix 2. The syntax examples and selected details have simply been updated to apply to the 2013-2015 NSFG.

To combine, pool, or "stack" data for male and female respondents in the 2013-2015 NSFG files, you subset the desired variables, and then append the 2 datasets. The CASEID values for males and females are non-overlapping, but you are advised to create a "sex of respondent" variable for use in your analyses. Assuming no subsetting of cases, the template programs below will append the female respondent file, which contains 5,699 records, to the male respondent file, which contains 4,506 records, into a SAS or Stata data set that contains 10,205 records.

Before pooling data for males and females, you may wish to consult **Appendix 4a,** the recode crosswalk showing comparable recodes for male and female respondents in the 2013-2015 NSFG. Note that any variable not found on the male or female data file will have all missing values on your combined data file. For example, CONSTAT1 is a recode only constructed for females, and if included for females, it will have all blank values for males in the combined data set. Before appending data, be sure to rename and/or redefine variables if variable names or response categories are different for males and females. Again, this information is included in the male-female recode crosswalk (Appendix 4a).

One additional note about sample design and weight variables:
These variables have the same names on both the male and female files for 2013-2015, and should require no renaming when you combine data from these files. See section on **SAMPLE WEIGHTS AND VARIANCE ESTIMATION** in Part 1 of User's Guide for further details.

- SEST
- SECU
- WGT2013_2015

Using SAS:

```
/* create subsets of male & female data and define R_SEX for sex of respondent */
      DATA FEMDATA;
      SET FEMRESP (KEEP=CASEID [other variables you wish to include]);
            R_SEX=1; /* female */
            RUN;

      DATA MALEDATA;
      SET MALERESP (KEEP=CASEID [other variables you wish to include]);
            R_SEX=2; /* male */
            RUN;

/* combined male & female data*/
      data MF_POOLED;
```

```
            set FEMDATA MALEDATA;

/* The SET statement will yield a combined file with 10,205 male & female records.
/* You could also obtain this result using the SAS Proc Append statement. */
```

Using Stata:

```
/* create subsets of male & female data and define R_SEX for sex of respondent */
      use FEMALE
      keep CASEID [other variables you wish to include]
      sort CASEID

/* create flag for female */
      generate R_SEX = 1
      save FEMSORTED, replace
      clear

      use MALE
      keep CASEID [other variables you wish to include]
      sort CASEID

/* create flag for male */
      generate R_SEX = 2
      save MALESORTED, replace

/* Append the 2 sorted files and create a variable to keep track of which
observations come from which dataset */
      append using FEMSORTED, gen(whichfile)
      save MALEFEMALE, replace

/* The above statements will yield a combined file with 10,205 male & female
records. */
```

Combining Data across NSFG Data File Releases

Note:
This section has been expanded and updated from the information provided in the 2011-2013 NSFG User's Guide Appendix 2.

Combining multiple NSFG data files may be beneficial for analyses that require a larger sample size; however, analysts should use caution in interpreting estimates based on data from the entire time period, as it may not be appropriate to estimate or interpret such "weighted averages" across broad spans of years. For example, if the estimates from separate data files vary significantly, the estimate derived from the combined data may be misleading. Also, the variations seen across the separate data files may be due to changes in the outcome of interest or changes in the population composition over time, or both. Similarly, one should be cautious when interpreting estimates from combined data files because the NSFG has not been conducted with a continuous nor annual survey design that would permit valid estimation and inference for the full span of years. The 1st six NSFG surveys were conducted using periodic survey design, with independent, nationally representative samples interviewed in 1973, 1976, 1982, 1988, 1995, and 2002. Even as the NSFG adopted a continuous interviewing design in 2006, there was a 15 month gap in interviewing from mid-June 2010 through mid-September 2011, such that data were not collected continuously from 2006-2015. Given this 15 month gap, when analyzing combined data from 2006-2010, 2011-2013, and 2013-2015 NSFG, *it is not appropriate to refer to these combined data as the 2006-2015 NSFG.*

Finally, while each year of NSFG fieldwork beginning in 2006 under the continuous design is based on an annual sample designed to be nationally representative, the sample sizes of interviews collected annually are not sufficient to provide statistically reliable estimates. For this reason, annual weights are not provided for the NSFG under the continuous fieldwork design begun in 2006. For the 2006-2010 and later files, the smallest period of estimation for which weights are provided is two years:

- WGTQ1Q8 and WGTQ9Q16 on the 2006-2010 files
- WGT2011_2013 on the 2011-2013 files
- WGT2013_2015 on the 2013-2015 files

The two-year sample weights contained in each NSFG data file for 2006-2010 NSFG and later represent the population at the approximate midpoint of data collection for that particular group of years (see table below). The weight for the 1995 NSFG file is adjusted to reflect population totals in May of 1995. The weight in the 2002 file reflects population totals in June of 2002.

Parameter estimates such as percentages and means for combined data files can be validly calculated and interpreted, albeit with appropriate caution if the survey periods combined are wide, using the code examples provided below. However, valid population sizes cannot be simply estimated based on combining multiple data files even when using the available 2-year or 4-year case weights. This is because each weight, whether for one of the periodically conducted NSFG surveys or for a 2-year or 4-year period of continuous fieldwork, is designed to represent the full US household population aged 15-44. For example, since the weights on each female file reflect the population size of roughly 61 million women aged 15-44 in the U.S. household population, if two female data files are combined with no further adjustment, the weights will result in a population size of roughly 122 million women 15-44.

Researchers may choose to scale down the weights (for example, dividing the weight value by two when combining two data files), however, the accuracy of the resulting population size estimates may still vary based on the span of years covered by the data and any population composition changes over that period of time, particularly with regard to age, race, and other factors for which the weights have been adjusted or post-stratified. For example, if the analyst combines 1995 and 2002 data, and divides the sample weights by 2, the resulting population size estimates will be appropriately scaled down but will not necessarily approximate the US household population at the midpoint of 1995 and 2002 because the weights were not designed to represent that point in time. Similarly, the user might combine data from 2006-2010 and 2011-2015 and scale down each of the 4-year file weights for those years by a factor of 2, but then the resulting population sizes and other parameter estimates cannot be reported as representing the whole span of time 2006-2015. The exception to this caveat is when combining 2011-2013 and 2013-2015 NSFG data, as described in the next paragraph.

Given the continuous fieldwork period and sample sizes associated with the separate 2-year files for 2011-2013 and 2013-2015, users are likely to combine these public use files to obtain sample sizes comparable to 2006-2010 and prior NSFG file releases. Therefore, for the full 2011-2015 NSFG survey period, a separate 4-year case weight has been provided **(WGT2011_2015)**, and is designed to represent population totals at the approximate midpoint of data collection (July 2013) over this four-year fieldwork period. *This weight variable **WGT2011_2015** should be used when users want to produce point estimates including population size estimates for the full 2011-2015 survey period based on combining the 2011-2013 and 2013-2015 data*. When using this 4-year weight to analyze 2011-2015 data, there is no need to scale down the resulting population sizes. If researchers are analyzing 2011-2015 data combined with earlier NSFG data files, the 2011-2015 file would count as 1 data file, not 2.

However, if users want to compare point estimates between the 2011-2013 and 2013-2015 survey periods, they should still use the 2-year weights for each separate file ***(WGT2011_2013 and WGT2013_2015)*** and create a variable to indicate each of the 2-year survey periods, or their midpoints of 2012 and 2014. (See example further below for each of these scenarios).

In summary:
When combining multiple NSFG data files, analysts should consider their specific analysis goals, define their populations carefully, and use caution when interpreting point estimates, particularly population size estimates, based on the combined data files.

The SAS and Stata syntax for combining data across NSFG data file releases is very similar to the syntax shown above for combining male and female data for 2013-2015. You subset the desired variables from each data set and then append or "stack" the 2 subsets. When selecting the variables for your analyses, you may wish to consult **Appendixes 4b and 4c**, which provide crosswalks of comparable recodes across female data for 2002, 2006-2010, 2011-2013, and 2013-2015 and across male data for 2002, 2006-2010, 2011-2013, and 2013-2015. You may also find helpful the summary of questionnaire changes made since the 2011-2013 NSFG (**Appendix 5**).

The main difference in the program syntax when combining data across NSFG data file releases is you must define new variables to hold the appropriate weight and sample design variable information because they may have different names on the separate files you are combining. (These variable names are shown in tables further below.) In the syntax examples below, these newly created variables are called WEIGHTVAR, STRATVAR, and PANELVAR. In addition to these new variables, your combined data file should include a categorical variable to indicate the data file or survey period in which the case was interviewed. This variable can be a simple categorical variable with values corresponding to each data file you wish to combine, such as '1995' or '2006-2010.' Or, as shown in the examples below, you might create a SURVEY variable with values based on the *year* represented by the weights for that particular file – for example, SURVEY=2008 for the 2006-2010 NSFG or SURVEY=2014 for the 2013-2015 NSFG.

A special note concerning sample design variables in 1995 NSFG:
For Cycle 5 (1995), a transformation of the stratum and cluster variables (COL_STR and PANEL) is required to ensure that there are no overlapping values with Cycle 6 (2002) or 2006-2010 female data. A suggested transformation is shown in the syntax examples below for combining 1995 NSFG data with later female data files. No transformation is needed to the sample design variables if you are only combining data from 2002 or later because the numbering for the primary sampling units did not overlap for these NSFG survey years.

A special note concerning CASEID in 2002 NSFG:
The 2002 SAS program statements create CASEID as an alphanumeric variable whereas CASEID is defined as a numeric variable in the SAS programs for 1995, 2006-2010, 2011-2013, and 2013-2015 NSFG. By removing the '$' that precedes the column location in the 2002 programs before creating a SAS system file, CASEID will be created as a numeric variable. As with any data manipulation programming, it is prudent to review the log files generated by your statistical software package to check for warnings and errors.

A special note concerning variable lengths when combining NSFG data files (updated November 2016):

Users may receive a warning message when combining data from multiple data files when using SAS, such as the one below:

```
WARNING: Multiple lengths were specified for the variable [variable name here] by
input data set(s). This can cause truncation of data.
```

This warning may occur because some variables may be all system-missing on one public use file but contain valid (non-sysmis) values on another public use file. When a variable is all sysmis on a file, the variable length is 1, which may be less than the length of the variable when it contains valid data. There are 3 typical scenarios where you may receive this warning about possible truncation of data when combining data files:

- For "enter all that apply" questions, one file being combined may have more mentions than the other file. For example, if one file has 3 mentions, and the other file has 4 mentions, the variable for the 4th mention will be all system-missing on the 1st file.

- For "loops" such as children, spouses, and partners, one file being combined may have cases with applicable data for more "loops" than the other file. For example, one file may have data for 3 former spouses, and the other file may have data for 4 former spouses, and in this scenario, the loop of variables for the 4th former spouse would be all sysmis in the 1st file.

- In some instances, including the special case of QUARTER described below, some variables were defined with different lengths (or numbers of columns) of different data files.

In these situations of varying variable lengths for the same variable, SAS will automatically default to the variable length specified in the first data set mentioned in the "set" or "merge" statement. The SAS log should be checked carefully when combining multiple NSFG files to prevent truncation of data. If any of these messages appear, the variable name will be noted. Users should adjust their code between the data and set statements to reflect the correct variable length of the variable in question. The next example shows how to do this specifically for the variable QUARTER.

A special note concerning QUARTER in 2011-2013 and 2013-2015 NSFG:

The QUARTER variable indicates the 12-week quarter in which the interview was conducted. Quarters 1-8 are contained on the 2011-2013 public use files, and quarters 9-16 are contained on the 2013-2015 NSFG. Due to the fact that QUARTER was defined as a 1 column variable in 2011-2013 and as a 2 column variable in 2013-2015, users must make some adjustment to their SAS or Stata coding to avoid truncating values on QUARTER when combining 2011-2013 and 2013-2015 NSFG data. In SAS, this code adjustment would be to add this length statement between the data and set statements when reading in the 2011-2013 data: `length quarter $2;`

Combining Data for Females: 1995, 2002, 2006-2010, 2011-2013, and 2013-2015

Below is a table showing the original sample design and weight variables in each female NSFG data file. This is followed by template programs in SAS and Stata, combining data for females.

(Note: The example program creates a combined file for all NSFG female data file releases from 1995 through 2013-2015 for illustrative purposes, however it is quite unlikely that any user would need to combine all of these data files. As described above, estimates of parameters, including population sizes, based on combined data must be conducted and interpreted with caution, as they do not represent the US household population at a clear point or period in time.)

Although not shown in the table, a 4-year weight **WGT2011_2015**, weighted to the July 2013 U.S. household population, is available to users in a separate file on the NSFG webpage for use when combining data from 2011-2013 and 2013-2015. This weight should be used when making estimates for the full 2011-2015 survey period, based on pooling 2011-2013 and 2013-2015 data. (See syntax example later in this document.) As noted above, if your goal is to study differences between 2011-2013 and 2013-2015, you should use the separate 2 year file weights from these 2 files, and create a survey period variable, possibly based on QUARTER, to indicate the survey period in which the interview occurred. Please note the special instruction above for redefining QUARTER as a 2-column variable in 2011-2013 NSFG in order to avoid truncation of this variable's values when combining with 2013-2015 NSFG data.

Design variable	**Cycle 5 (1995) N=10,847**	**Cycle 6 (2002) N=7,643**	**2006-2010 N=12,279**	**2011-2013 N=5,601**	**2013-2015 N=5,699**
Stratum variable	COL_STR	SEST	SEST	SEST	SEST
Cluster/Panel variable	PANEL	SECU_R – fem resp SECU_P – fem preg	SECU	SECU	SECU
Final post-stratified, fully adjusted case weight	POST_WT	FINALWGT	WGTQ1Q16	WGT2011_2013	WGT2013_2015
What point in time does sample weight represent?	May 1995	June 2002	June 2008	July 2012	July 2014

Using SAS:

```
DATA NSFG95;
      set FEM95 (keep=caseid col_str panel post_wt [variables from female Cycle 5
                respondent file] );
      STRATVAR=COL_STR*200; /* just an example - other transformations possible */
      PANELVAR=PANEL*200;
      WEIGHTVAR=POST_WT;
      drop col_str panel post_wt;
      SURVEY=1995; /* value to indicate survey year */
      run;
```

```
DATA NSFG02;
      set FEM02 (keep=caseid sest secu_r finalwgt [variables from female Cycle 6
                respondent file]);
      STRATVAR=SEST;
      PANELVAR=SECU_R;
      WEIGHTVAR=FINALWGT;
      drop sest secu_r finalwgt;
      SURVEY=2002; /* value to indicate survey year */
      run;

DATA NSFG0610;
      SET fem0610 (keep=caseid sest secu wgtq1q16 [variables from female NSFG 2006-
                  2010 respondent file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR=WGTQ1Q16;
      drop sest secu wgtq1q16;
      SURVEY=2008; /* 2008 chosen due to year represented by the weight, but can
                      label as '2006-2010' instead */
      run;

DATA NSFG1113;
      SET fem1113 (keep=caseid sest secu wgt2011_2013 [additional variables from
                  female NSFG 2011-2013 respondent file]);
      /* add length statement for QUARTER variable if needed for your analysis */
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR=WGT2011_2013;
      drop sest secu wgt2011_2013;
      SURVEY=2012; /* 2012 chosen due to year represented by the weight, but can
                      label as '2011-2013' instead */
      run;

DATA NSFG1315;
      SET fem1315 (keep=caseid sest secu wgt2013_2015 [additional variables from
                  female NSFG 2013-2015 respondent file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR=WGT2013_2015;
      drop sest secu wgt2013_2015;
      SURVEY=2014; /* 2014 chosen due to year represented by the weight, but can
                      label as '2013-2015' instead */
      run;

DATA ALLFEMALE; SET NSFG95 NSFG02 NSFG0610 NSFG1113 NSFG1315;
RUN;
```

Using Stata:

```
use c5FemResp
keep CASEID COL_STR PANEL POST_WT (variables from female Cycle 5 respondent file)
gen STRATVAR=COL_STR*200  /* just an example - other transformations possible */
gen PANELVAR=PANEL*200
gen WEIGHTVAR=POST_WT
gen SURVEY=1995 /* value to indicate survey year */
drop COL_STR PANEL POST_WT
save c5Femnew, replace
CLEAR

use c6FemResp
```

```
keep CASEID SEST SECU_R FINALWGT (variables from female Cycle 6 respondent file)
gen STRATVAR=SEST
gen PANELVAR=SECU_R
gen WEIGHTVAR=FINALWGT
gen SURVEY=2002 /* value to indicate survey year */
drop SEST SECU_R FINALWGT
save c6Femnew, replace
CLEAR

use femresp0610
keep caseid sest secu wgtq1q16(variables from female NSFG 2006-2010 respondent
                               file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= wgtq1q16
gen SURVEY=2008               /* 2008 chosen due to year represented by the weight,
drop SEST SECU wgtq1q16       but can label as '2006-2010' instead */
save Femnew0610, replace


use femresp1113
keep caseid sest secu wgt2011_2013 (variables from female NSFG 2011-2013 respondent
                                    file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= wgt2011_2013
gen SURVEY=2012               /* 2012 chosen due to year represented by the weight,
drop SEST SECU wgt2011_2013   label as '2011-2013' instead */
save Femnew1113, replace

use femresp1315
keep caseid sest secu wgt2013_2015 (variables from female NSFG 2013-2015 respondent
                                    file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= wgt2013_2015
gen SURVEY=2014               /* 2014 chosen due to year represented by the weight,
drop SEST SECU wgt2013_2015   label as '2013-2015' instead */
save Femnew1315, replace

/*Append the 2011-2013 records to the end of the NSFG 2013-2015 data set*/
append using Femnew1113

/*Append the 2006-2010 records to the end of the combined records from NSFG 2013-
2015 and NSFG 2011-2013 data sets*/
append using Femnew0610

/*Append the Cycle 6 records to the end of the combined records from the NSFG 2013-
2015, NSFG 2011-2013 and NSFG 2006-2010 data sets*/
append using C6Femnew

/*Append Cycle 5 records to the end of the combined records from the NSFG 2013-
2015, NSFG 2011-2013, NSFG 2006-2010 and Cycle 6 data sets*/
append using C5FemNEW

/*Create permanent data file with concatenated records from the 5 data sets*/
save ALLFEMALE, replace
```

Combining Data for Males: 2002, 2006-2010, 2011-2013, and 2013-2015

Below is a table showing the original sample design and weight variables in each male NSFG data file. This is followed by template programs in SAS and Stata, combining data for males.

(Note: The example program creates a combined file for all NSFG male data file releases from 2002 through 2013-2015 for illustrative purposes, however it is quite unlikely that any user would need to combine all of these data files. As described above, estimates of parameters, including population sizes, based on combined data must be conducted and interpreted with caution, as they do not represent the US household population at a clear point or period in time.)

Although not shown in the table, a 4-year weight **WGT2011_2015**, weighted to the July 2013 U.S. household population, is available to users in a separate file on the NSFG webpage for use when combining data from 2011-2013 and 2013-2015. This weight should be used when making estimates for the full 2011-2015 survey period, based on pooling 2011-2013 and 2013-2015 data. (See syntax example later in this document.) As noted above, if your goal is to study differences between 2011-2013 and 2013-2015, you should use the separate 2 year file weights from these 2 files, and create a survey period variable, possibly based on QUARTER, to indicate the survey period in which the interview occurred. Please note the special instruction above for redefining QUARTER as a 2-column variable in 2011-2013 NSFG in order to avoid truncation of this variable's values when combining with 2013-2015 NSFG data.

Design variable	**Cycle 6 (2002) N=4,928**	**2006-2010 N=10,403**	**2011-2013 N=4,815**	**2013-2015 N=4,506**
Stratum variable	SEST	SEST	SEST	SEST
Cluster/Panel Variable	SECU	SECU	SECU	SECU
Final post-stratified, fully adjusted case weight	FINALWGT	WGTQ1Q16	WGT2011_2013	WGT2013_2015
What point in time does the sample weight represent?	June 2002	June 2008	July 2012	July 2014

Using SAS:

```
DATA NSFG02;
      set Male02 (keep=caseid sest secu finalwgt
            [variables from male Cycle 6 file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR=FINALWGT;
      drop sest secu finalwgt;
      SURVEY=2002;   /* value to indicate survey year */
      run;

DATA NSFG0610;
```

```
      SET MALE610 (keep=caseid sest secu WGTQ1Q16
      [variables from male 2006-2010 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR= WGTQ1Q16;
      drop sest secu WGTQ1Q16;
      SURVEY=2008; /* 2008 chosen due to year represented by the weight, but can
                      label as '2006-2010' instead */
      run;

DATA NSFG1113;
      SET MALE1113 (keep=caseid sest secu WGT2011_2013
      [variables from male 2011-2013 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR= WGT2011_2013;
      drop sest secu WGT2011_2013;
      SURVEY=2012; /* 2012 chosen due to year represented by the weight, but can
                      label as '2011-2013' instead */
      run;

DATA NSFG1315;
      SET MALE1315 (keep=caseid sest secu WGT2013_2015
      [variables from male 2013-2015 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR= WGT2013_2015;
      drop sest secu WGT2013_2015;
      SURVEY=2014; /* 2014 chosen due to year represented by the weight, but can
                      label as '2013-2015' instead */
      run;

DATA ALLMALE; SET NSFG02 NSFG0610 NSFG1113 NSFG1315;
RUN;
```

Using Stata:

```
use C6MALERESP
keep CASEID SECU SEST FINALWGT (variables from male Cycle 6 file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR=FINALWGT
gen SURVEY=2002  /* value to indicate survey year */
save C6MALENEW, replace
clear

use MALERESP0610
keep CASEID SEST SECU WGTQ1Q16 (variables from male 2006-2010 NSFG file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= WGTQ1Q16
gen SURVEY=2008         /* 2008 chosen due to year represented by the weight, but
                           can label as '2006-2010' instead */
save MALENEW0610, replace

use MALERESP1113
keep CASEID SEST SECU WGT2011_2013 (variables from male 2011-2013 NSFG file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= WGT2011_2013
```

```
gen SURVEY=2012          /* 2012 chosen due to year represented by the weight, but
                         can label as '2011-2013' instead */
save MALENEW1113, replace

use MALERESP1315
keep CASEID SEST SECU WGT2013_2015(variables from male 2013-2015 NSFG file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= WGT2013_2015
gen SURVEY=2014          /* 2014 chosen due to year represented by the weight, but
                         can label as '2013-2015' instead */
save MALENEW1315, replace

/*Append the NSFG 2011-2013 records to the end of the NSFG 2013-2015 data set*/
append using MALENEW1113

/*Append the NSFG 2006-2010 records to the end of the combined records from the
NSFG 2011-2013 and 2013-2015 data sets*/
append using MALERESP0610

/*Append the Cycle 6 records to the end of the combined records from the NSFG 2006-
2010, 2011-2013 and 2013-2015 data sets*/
append using C6MALENEW

/*Create permanent data file with concatenated records from the 4 data sets*/
save ALLMALE, replace
```

Creating a 2011-2015 NSFG Data File for Analysis, or Combining with Earlier File Releases

The syntax examples above refer to the separate 2-year file weights for 2011-2013 and 2013-2015 and should be used if the goal is to compare estimates *between these 2 survey periods.* If the user instead wishes to combine these two 2-year files and do analyses based on the full 4 years of data for 2011-2015, possibly also comparing 2011-2015 data to earlier NSFG survey periods, the 4-year weight WGT2011_2015 should be used.

The SAS and Stata syntax below show one example of combining 2006-2010 and 2011-2015 NSFG male data for such analyses. Similar syntax would be used to combine female data for these survey periods. The 1st step in these program statements is to combine 2011-2013 and 2013-2015 data and merge in the 4 year weight variable accessible separately on the NSFG webpage. This portion of the example may be helpful to users who simply want to analyze 2011-2015 NSFG data as a single survey period.

Note that if population sizes are also to be estimated for the combined file with 2006-2010 and 2011-2015 data, the weights in these examples must be scaled down by a factor of 2 because 2 data files are being combined, each with weights that would otherwise scale to the full population. However, the survey period that is represented cannot be reported as 2006-2015, primarily due to the 15 month gap in interviewing between those survey periods, but also because the weights were not designed or adjusted for this purpose.

Using SAS:

```
/* first combine 2011-2013 and 2013-2015 data files into a single 2011-2015 file*/
DATA NSFG1113;
      SET MALE1113 (keep=caseid sest secu
      [variables from male 2011-2013 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      drop sest secu;
```

```
      run;

DATA NSFG1315;
      SET MALE1315 (keep=caseid sest secu
      [variables from male 2013-2015 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      drop sest secu;
      RUN;

DATA NSFG1115;
      SET NSFG1113 NSFG1315;
      SURVEY=2013; /* 2013 chosen due to year represented by the 4 year weight,
                      but can label as '2011-2015' instead */
RUN;
PROC SORT; BY CASEID;
RUN;

DATA FOURYR_WEIGHT;
      SET WGT1115; /* read in sasfile with 4 year weight for males */
RUN;
PROC SORT; BY CASEID;
RUN;

/* merge 4 year weight onto 2011-2015 combined file, based on CASEID */
DATA MALE1115;
      MERGE NSFG1115 (IN=A) FOURYR_WEIGHT; BY CASEID; IF A;
      WEIGHTVAR=WGT2011_2015;
      Drop WGT2011_2015;
      SURVEY=2013; /* 2013 chosen due to year represented by the weight, but can
                      label as '2011-2015' instead */
      run;

/* only need the code above if only interested in analyzing 2011-2015 data */

/* Now read in 2006-2010 NSFG male data */
DATA MALE0610;
      SET 0610MALE (keep=caseid sest secu WGTQ1Q16
      [variables from male 2006-2010 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR= WGTQ1Q16;
      drop sest secu WGTQ1Q16;
      SURVEY=2008; /* 2008 chosen due to year represented by the weight, but can
                      label as '2006-2010' instead */
      run;

/* Combining male data from 2006-2010 and 2011-2015 */
DATA ALLMALE; SET MALE0610 MALE1115;
RUN;

/* The ALLMALE file contains WEIGHTVAR based on the 2 separate 4-year file weights
   from each data file 2006-2010 and 2011-2015.  The SURVEY variable permits
   testing of differences in estimates between these two 4-year survey periods. */

/* If the user wishes to estimate population sizes based on the combined "ALLMALE"
   data file, WEIGHTVAR should be scaled down by a factor of 2. */
```

Using Stata:

```
/* first combine 2011-2013 and 2013-2015 data files into a single 2011-2015 file*/
```

```
      use MALERESP1113
      keep CASEID SEST SECU (variables from male 2011-2013 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      save MALENEW1113, replace

      use MALERESP1315
      keep CASEID SEST SECU (variables from male 2013-2015 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      save MALENEW1315, replace

/*Append the NSFG 2011-2013 records to the end of the NSFG 2013-2015 data set*/
      append using MALENEW1113
      gen SURVEY=2013 /* 2013 chosen due to year represented by 4 year weight,
                         but can label as '2011-2015' instead */
      sort CASEID
      save MALE1115, replace

      use WGT1115 /* read in file with 4 year weights for males & sort by merge
                     variable*/
      sort CASEID
      save FOURYR_WEIGHT, replace

      /* merge 4 year weight onto 2011-2015 combined file, based on CASEID */
      merge 1:1 CASEID using MALE1115
      gen WEIGHTVAR=WGT2011_2015
      keep if _merge==3
      save MALE1115, replace

      /* only need the code above if only interested in analyzing 2011-2015 data */

      /* Now read in 2006-2010 NSFG male data */
      use MALERESP0610
      keep CASEID SEST SECU WGTQ1Q16 (variables from male 2006-2010 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      gen WEIGHTVAR= WGTQ1Q16
      gen SURVEY=2008 /* 2008 chosen due to year represented by the weight, but can
                         label as '2006-2010' instead */
      save MALENEW0610, replace

      /* Combining male data from 2006-2010 and 2011-2015 */
      append using MALE1115
      save ALLMALE, replace

      /* The ALLMALE file contains WEIGHTVAR based on the 2 separate 4-year file weights
         from each data file 2006-2010 and 2011-2015.  The SURVEY variable permits
         testing of differences in estimates between these two 4-year survey periods. */

      /* If the user wishes to estimate population sizes based on the combined "ALLMALE"
         data file, WEIGHTVAR should be scaled down by a factor of 2. */
```