2015-2017 NSFG USER'S GUIDE APPENDIX 2:
SAS AND STATA SYNTAX GUIDELINES
FOR COMBINING DATA ACROSS FILE RELEASES

This appendix provides technical guidance, including suggested SAS and Stata syntax, for combining NSFG data across file releases.

These guidelines and examples assume 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 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 on the NSFG webpage.

Combining Data across NSFG File Releases

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 (particularly with age, race, and other factors for which the weights have been adjusted or post-stratified), or both. 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 with continuous interviewing, adopted by the NSFG beginning in 2006, breaks in data collection occur due to the need to award a new contract when an existing one ends. Thus, there was a 15-month gap in interviewing from mid-June 2010 through mid-September 2011, such that data were not collected continuously from 2006-2017. **Given this 15-month gap, when analyzing combined data from 2006-2010, 2011-2013, 2013-2015 and 2015-2017 NSFG, *it is not appropriate to refer to these combined data as the 2006-2017 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. In addition, the weights and the design variables (SEST and SECU) are not designed to produce yearly estimates. For these reasons, 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
- WGT2015_2017 on the 2015-2017 files

Given the continuous fieldwork period and sample sizes associated with the separate two-year files for 2011-2013, 2013-2015, and 2015-2017, separate four-year and six-year case weights have been provided**:**

- WGT2011_2015
- WGT2013_2017
- WGT2011_2017

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. However, valid population sizes cannot be estimated based on combining multiple data files even when using the available two-year, four-year, or six-year case weights. This is because each weight, whether for one of the periodically conducted NSFG surveys or for a two-year, four-year or six-year period of continuous interviewing, is designed to represent the full U.S. household population aged 15-44 (and up to age 49 in 2015-2017) at the approximate midpoint of data collection. For example, the weights on each female file reflect the population size of roughly 61 million women aged 15-44 in the U.S. household population, thus 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. Therefore, 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 time period.

The separate four-year (WGT2011_2015 and WGT2013_2017) and six-year (WGT2011_2017) case weights are designed to represent population totals for men and women aged 15-44 at the approximate midpoint of data collection (July 2013, July 2015, and July 2014, respectively) over these fieldwork periods. *They should be used when users want to produce point estimates including population size estimates for any of these survey periods, 2011-2015, 2013-2017 or 2011-2017, based on combining the 2011-2013, 2013-2015 and 2015-2017 data*. This approach of using the prepared four-year or six-year case weights essentially treats these combined survey periods as a single survey period for purposes of statistical inference, and users should consider carefully whether this approach is best suited for their particular analysis.

When using these weights to analyze NSFG data, there is no need to scale down the resulting population sizes. But if researchers want to combine recent data files with 2006-2010 or older data and older, note that no separate sample weights are provided, and users must scale down weights if interested in making estimates of population sizes. If researchers are analyzing 2011-2015 data combined with earlier NSFG data files, the 2011-2015 file would count as one data file, not two. *However, if users want to compare point estimates between the 2011-2013, 2013-2015, and 2015-2017 survey periods, they should still use the two-year weights for each of the separate files (**WGT2011_2013, WGT2013_2015, and WGT2015_2017**)* and create a variable to indicate each of the two-year survey periods, or their midpoints of 2012, 2014, and 2016.

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

The SAS and Stata syntax for combining NSFG data across file releases requires that 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 and male data for 2006-2010, 2011-2013, 2013-2015 and 2015-2017. You may also find helpful the summary of questionnaire changes made since the 2013-2015 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. (See table 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 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 '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=2016 for the 2015-2017 NSFG.

A special note concerning variable lengths when combining NSFG data files:

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 three mentions, and the other file has four mentions, the variable for the fourth mention will be all system-missing on the first 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 three former spouses, and the other file may have data for four former spouses, and in this scenario, the loop of variables for the fourth former spouse would be all sysmis in the firstfile.

- In some instances, including the special case of QUARTER described below, some variables were defined with different lengths (or numbers of columns) in 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, 2013-2015, and 2015-2017 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, quarters 9-16 are contained on the 2013-2015 public use file, and quarters 17-24 are contained in the 2015-2017 NSFG. Due to the fact that QUARTER was defined as a one-column variable in 2011-2013 and as a two-column variable in 2013-

2015 and 2015-2017, users must make some adjustment to their SAS or Stata coding to avoid truncating values on QUARTER when combining 2011-2013 with the2013-2015 and 2015-2017 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 from 2006-2010, 2011-2013, 2013-2015, and 2015-2017

Below is a table showing the original sample design and weight variables in each NSFG file release since 2006-2010.

Design variable	2006-2010	2011-2013	2013-2015	2015-2017
Stratum variable	SEST	SEST	SEST	SEST
Cluster/Panel variable	SECU	SECU	SECU	SECU
Final post-stratified, fully adjusted case weight	WGTQ1Q16	WGT2011_2013	WGT2013_2015	WGT2015_2017
What point in time does sample weight represent?	June 2008	July 2012	July 2014	July 2016

Note: See the 2013-2015 User's Guide Appendix 2 for more information about 1995 and 2002 sample design and weight variables.

As noted above, if your goal is to study differences between 2011-2013 and 2013-2015 and 2015-2017, you should use the separate two-year file weights from these three 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 two-column variable in 2011-2013 NSFG in order to avoid truncation of this variable's values when combining with 2013-2015 and/or 2015-2017 NSFG data.

Keep in mind that beginning in September 2015, the NSFG expanded its age range from 15-44 to 15-49. Users need to be cautious when combining the 2015-2017 NSFG with older data files because respondents aged 45-49 will not have values on the four-year and six-year case weights representing the survey periods of 2013-2017 and2011-2017. Below are the sample sizes you will have for respondents aged 15-44 when combining 2015-2017 data with 2011-2013 and 2013-2015 data:

Combining data for:	Total N	Males	Females	Overall Response Rate	Male Response Rate	Female Response Rate
2013-2017	19,095	8,505	10,590	67.4%	65.3%	69.1%
2011-2017	29,511	13,320	16,191	69.2%	67.6%	70.6%

In addition, if your goal is to study differences between file releases, you should conduct your analysis using the subpopulation of respondents aged 15-44, using appropriate survey software commands for indicating that

subpopulation.

Below are template programs in SAS and Stata, combining data for women. This example uses female data but similar syntax should be used when combining data for men across survey periods.

Using SAS:

```
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 NSFG1517;
      SET fem1517 (keep=caseid sest secu wgt2015_2017 [additional variables from
                  female NSFG 2015-2017 respondent file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      WEIGHTVAR=WGT2015_2017;
      drop sest secu wgt2015_2017;
      SURVEY=2016; /* 2016 chosen due to year represented by the weight, but can
                      label as '2015-2017' instead */
      run;

DATA ALLFEMALE; SET NSFG0610 NSFG1113 NSFG1315 NSFG1517;
RUN;
```

Using Stata:

```
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

use femresp1517
keep caseid sest secu wgt2015_2017 (variables from female NSFG 2015-2017 respondent
                                    file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR=wgt2015_2017
gen SURVEY=2016                /* 2016 chosen due to year represented by the weight,
drop SEST SECU wgt2015_2017    label as '2015-2017' instead */
save Femnew1517, replace

/*Append the 2013-2015 records to the end of the NSFG 2015-2017 data set*/
append using Femnew1315

/*Append the 2011-2013 records to the end of the combined NSFG 2013-2015 and NSFG
2015-2017 data sets*/
append using Femnew1113

/*Append the 2006-2010 records to the end of the combined records from NSFG 2011-
2013, NSFG 2013-2015, and NSFG 2015-2017 sets*/
append using Femnew0610


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

Creating a 2011-2017 Combined File and Using the Prepared Six-Year Weight

The SAS and Stata syntax below show one example of combining 2011-2013, 2013-2015 and 2015-2017 NSFG male data. Similar syntax could be used to combine female data for this six-year period or for combining 2013-2015 and 2015-2017 into a four-year file using the prepared four-year weight WGT2013_2017. However, for illustrative purposes, here we show how to combine all three files into a six-year file and make use of the prepared six-year weight WGT2011_2017.

The first step in these program statements below is to combine the separate two-year data sets into one six-year file. The second step is to merge in the six-year weight variable accessible separately on the NSFG webpage.

As noted above, due to the age range expansion of the NSFG to 15-49 in September 2015, the separate four-year and six-year case weights only have values for respondents aged 15-44. Respondents aged 45-49 in 2015-2017 will be missing values on those weights, and users must bear this in mind when combining the 2015-2017 NSFG data with older file releases.

Using SAS:

```
/* first combine 2011-2013, 2013-2015 and 2015-2017 data files into a single 2011-
2017 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 NSFG1517;
      SET MALE1517 (keep=caseid sest secu
      [variables from male 2015-2017 NSFG file]);
      STRATVAR=SEST;
      PANELVAR=SECU;
      drop sest secu;
      RUN;


DATA NSFG1117;
      SET NSFG1113 NSFG1315 NSFG1517;
      SURVEY=2014; /* 2014 chosen due to year represented by the six-year weight,
                      but can label as '2011-2017' instead */
RUN;
PROC SORT; BY CASEID;
RUN;

DATA SIXWEIGHT;
      SET MALEWGT1117 (keep= caseid wgt2011_2017); /* read in sasfile with six-year
weight for men */
```

```
RUN;
PROC SORT; BY CASEID;
RUN;

/* merge six-year weight onto 2011-2017 combined file, based on CASEID */
DATA MALE1117;
      MERGE NSFG1117 (IN=A) SIXWEIGHT; BY CASEID; IF A;
      WEIGHTVAR=WGT2011_2017;
      Drop WGT2011_2017;

      run;
```

Using Stata:

```
/* first combine 2011-2013, 2013-2015, and 2015-2017 data files into a single 2011-2017
file*/
      use MALERESP1113
      keep CASEID SEST SECU (variables from male 2011-2013 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      drop SEST SECU wgt2011_2013
      save MALENEW1113, replace

      use MALERESP1315
      keep CASEID SEST SECU (variables from male 2013-2015 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      drop SEST SECU wgt2013_2015
      save MALENEW1315, replace

      use MALERESP1517
      keep CASEID SEST SECU (variables from male 2015-2017 NSFG file)
      gen STRATVAR=SEST
      gen PANELVAR=SECU
      drop SEST SECU wgt2015_2017
      save MALENEW1517, replace

/*Append the 2013-2015 records to the end of the NSFG 2015-2017 data set*/
      append using MALENEW1315
      sort CASEID

/*Append the 2011-2013 records to the end of the combined NSFG 2013-2015 and NSFG 2015-
2017 data sets*/
      append using MALENEW1113

      sort CASEID
      save MALE1117, replace

      use MALEWGT1117 /* read in file with six-year weights for men & sort by merge
                      variable*/
      keep CASEID wgt2011_2017
      sort CASEID
      save SIXYR_WEIGHT, replace

      /* merge six-year weight onto 2011-2017 combined file, based on CASEID */
      merge 1:1 CASEID using MALE1117
      gen WEIGHTVAR=WGT2011_2017
      drop wgt2011_2017
      keep if _merge==3
      save MALE1117, replace
```