2006-2010 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 3 types of file manipulations that are commonly done with the NSFG public use data.

- Combining 2006-2010 data from female respondent and female pregnancy files
- Combining 2006-2010 data from male and female files
- Combining data across years of data collection –
    - 1995, 2002, and 2006-2010 for females
    - 2002 and 2006-2010 for males

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 selected variance estimation scenarios are posted on the NSFG webpage.

Combining Data from Female Respondent and Female Pregnancy Files:

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 19 pregnancies. (No respondent in the 2006-2010 NSFG reported more than 19 pregnancies, though space was allowed for 20.) See **Appendix 1 (File Indexes)** for more details and file location.

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

-- pregnancy outcome (OUTCOM01-19)
-- date pregnancy ended (DATEND01-19)
-- age of woman at time of pregnancy outcome (AGEPRG01-19)
-- formal marital status at pregnancy outcome (MAROUT01-19)
-- informal marital status at pregnancy outcome (RMAROUT01-19)
-- date of conception (DATCON01-19)
-- age of women at time of conception (AGECON01-19)
-- formal marital status at conception (MARCON01-19)

-- informal marital status at conception (RMARCON01-19)
-- current living situation of 1st liveborn child from the pregnancy (LIVCHILD01-19)
-- wantedness of pregnancy by R at time of conception (Cycle 4 definition) (OLDWR01-19)
-- wantedness of pregnancy by H/P at time of conception (Cycle 4 definition) (OLDWP01-19)
-- wantedness of pregnancy by R at time of conception (Cycle 6 definition) (WANTRP01-19)
-- wantedness of pregnancy by H/P at time of conception (Cycle 6 definition) (WANTP01-19)

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 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 12,279 records) or pregnancy (interval)-based (up to 20,492 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
- WGTQ1Q16 (for use with all 16 quarters of data)
- FINALWGT30 (for use when analyzing the 1st 2 1/2 years of data collection - quarters 1 to 10)
- WGTQ9Q16 (for use when analyzing years 3 and 4 of data collection - quarters 9 to 16)
- WGTQ5Q16 (for use when analyzing years 2 - 4 of data collection - quarters 5 to 16)
- WGTQ1Q8 (for use when analyzing years 1 and 2 of data collection - quarters 1 to 8)

Adding Respondent Variables to a Pregnancy (Interval) Based File

Using SAS:

This template program will yield a SAS data file with a maximum of 20,492 *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 use a line such as the one shown in red below.*

```
/*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 2006-2010 sasfile with:
      20,492 records if the line in red is NOT used, and
      14,292 records if the line in red is used to subset only live births */
```

Using Stata:

This template program will yield a Stata file with a maximum of 20,492 *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 use a line such as the one shown in red below.*

```
* 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]
* this line subsets only those pregnancies ending in live birth
KEEP IF OUTCOME==1
SORT CASEID
SAVE PREGSORTED, REPLACE

* Merge the 2 sorted files
MERGE m:1 CASEID USING RESPONDSORTED
KEEP IF _MERGE==3

* The above merge will produce a 2006-10 .dta file with:
*     20,492 records if the line in red is NOT used, and
*     14,292 records if the line in red is used to subset only live births
```

Adding Pregnancy Variables to a Respondent Based File

The SAS & Stata templates 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 12,279 women, not all of them have ever been pregnant, so the maximum records you could have based on ever-pregnant women is 7,538 (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 most recent live birth for each respondent . Such a subsetted file may be helpful if you wish to examine, for example, breastfeeding and maternity leave for the most recent birth in the context of the respondent's contraceptive, work, or relationship experiences.*

For this program, the following pregnancy variables are needed:

CASEID = Case identification number
OUTCOME = Outcome of pregnancy (=1 if live birth)
PREGORDR = Pregnancy order or number

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; /* this line subsets only those pregnancies ending in live
                  birth */
RUN;

/*Sort 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 2006-2010 sasfile with 6,683 records, which
is consistent with the respondent-based recode, LBPREGS, which indicates how
many pregnancies resulted in live birth for each respondent (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 2006-2010 .dta file with 6,683 records,
which is consistent with the respondent-based recode, LBPREGS, which
indicates how many pregnancies resulted in live birth for each respondent
(see Appendix 3a).
```

Combining Data from Male and Female Files From 2006-2010

To combine or pool data for males and females in the 2006-2010 NSFG files, you subset the desired variables, and then append the 2 data sets. 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 12,279records, to the male respondent file, which contains 10,403records, into a SAS or Stata data set that contains 22,682records

Before pooling data for males and females, you may wish to consult **Appendix 4a,** the recode crosswalk showing comparable recodes by gender in the 2006-2010 NSFG. Note that any variable not found on the male or female data file will have all missing values for the records from the data file from which it was missing when the data files are combined. 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, type or coding is different for males and females.

One additional note about sample design and weight variables: These variables have the same names on both the male and female files for 2006-2010, 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.

- WGTQ1Q16
- SEST
- SECU

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;
data MF_POOLED;
      set femdata maledata;
/* The SET statement will yield a combined M-F dataset with 22,682 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 M-F dataset with 22,682 records.
```

Combining Data Across Years of NSFG Data

The SAS and Stata syntax for combining data across years of NSFG is very similar to the syntax shown above for combining male and female data in 2006-2010. You subset the desired variables from each data set and then append. 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 1995, 2002, and 2006-2010, and across male data for 2002 and 2006-2010. You may also find helpful the summary of questionnaire changes from 2002 through 2006-2010 NSFG (**Appendix 5**).

The main difference in the program syntax when combining data across NSFG years is you must define new variables to hold the appropriate weight and sample design variables. Note that 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. There is no need to do this for 2002 and 2006-2010 because the primary sampling units do not overlap for these survey years.

Additionally, the 2002 SAS program statements create CASEID as an alphanumeric variable whereas CASEID is numeric in the 1995 and 2006-2010 programs. 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 SAS to check for warnings and errors.

Combining Data for Females: 1995, 2002, 2006-2010

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.

Design variable	**Cycle 5 (1995)**	**Cycle 6 (2002)**	**2006- 2010**
Stratum variable	COL_STR	SEST	SEST
Cluster/Panel Variable	PANEL	SECU_R – fem resp SECU_P – fem preg	SECU
Final post-stratified, fully adjusted case weight	POST_WT	FINALWGT	WGTQ1Q16

Using SAS:

```
DATA NSFG95;
      set FEM95 (keep=caseid col_str panel post_wt
            [variables from female Cycle 5 respondent file] );
      STRATVAR=COL_STR+200; /* other transformations possible,
                               but this one works well */
      PANELVAR=PANEL+200;
      WEIGHTVAR=POST_WT;
      drop col_str panel post_wt;
      SURVEY=1995;  /* use some 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;
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=2006;
run;

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

Using Stata:

```
use c5FemResp
keep CASEID COL_STR PANEL POST_WT
            (variables from female Cycle 5 respondent file)
gen STRATVAR=COL_STR*200  /* other transformations possible,
                              but this one works well */
gen PANELVAR=PANEL*200
gen WEIGHTVAR=POST_WT
gen SURVEY=1995  /* use some 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
drop SEST SECU_R FINALWGT
save c6Femnew, replace
CLEAR

use c7femresp
keep caseid sest secu wgtq1q16
            (variables from female NSFG 2006-2010 respondent file)
gen STRATVAR=SEST
gen PANELVAR=SECU
gen WEIGHTVAR= wgtq1q16
gen SURVEY=2006
drop SEST SECU wgtq1q16
save c7Femnew, replace

/*Append the Cycle 6 records to the end of the NSFG 2006-2010 data set*/
append using C6FemNEW
```

```
/*Append Cycle 5 records to the end of the combined records from 2006-2010
and Cycle 6*/
append using C5FemNEW

/*Create permanent data file with concatenated records from the 3 data sets*/
save ALLFEMALE, replace
```

Combining Data for Males: 2002 and 2006-2010

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.

Design variable	**Cycle 6 (2002) N=4,928**	**2006-2010 N=10,403**
Stratum variable	SEST	SEST
Cluster/Panel Variable	SECU	SECU
Final post-stratified, fully adjusted case weight	FINALWGT	WGTQ1Q16

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;   /* use some 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=2006;
run;

DATA ALLMALE; SET NSFG02 NSFG0610;
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  /* use some value to indicate survey year */
save C6MALENEW, replace
clear

use C7MALERESP
keep caseid sest secu WGTQ1Q16
            (variables from male 2006-2010 NSFG file)
gen STRATVAR=sest
gen PANELVAR=secu
gen WEIGHTVAR= WGTQ1Q16
gen SURVEY=2006
save C7MALENEW, replace

/*Append the Cycle 6 records to the end of the NSFG 2006-2010 data set*/
append using C6MALENEW

/*Create permanent data file with concatenated records from the 2 data sets*/
save ALLMALE, replace
```