State Youth Tobacco Survey (YTS) Methodology Report

Prepared for

Office on Smoking and Health
Epidemiology Branch
CDC/ONDIEH/NCCDPHP/OSH/EB
4770 Buford Highway, N.E., Mailstop F79
Atlanta, GA 30341

Prepared by

RTI International
3040 E. Cornwallis Road
Research Triangle Park, NC 27709

April 2018

Recommended citation:

Office on Smoking and Health. State Youth Tobacco Survey (YTS) Methodology Report. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Health Promotion, Office on Smoking and Health, 2018.

CONTENTS

1. INTRODUCTION

1.1 Purpose

This document contains a description of the Youth Tobacco Survey (YTS) data processing, weighting, and analysis procedures.

1.2 Background

The Youth Tobacco Survey (YTS), conducted by State Departments of Health, collects data in school settings from students in grades 6 through 12. Analysis of YTS data measures the prevalence of tobacco products (e.g., cigarettes, cigars, bidis, tobacco pipes, smokeless tobacco, snus, dissolvable tobacco products, hookahs, and electronic cigarettes), knowledge and attitudes regarding tobacco use, exposure to pro-tobacco media and advertising, information on the enforcement of prohibiting the sale of tobacco to minors, presence of tobacco control programs in school curricula, cessation attempts and successes, and second-hand smoke exposure. The YTS is intended to enhance the capacity of state agencies and organizations to design, implement, and evaluate tobacco prevention and control programs for the purposes of preventing U.S. youths from using tobacco products and helping current users to quit.

While individual State Departments of Health plan and implement the YTS survey, the Youth Tobacco Survey Handbook, disseminated by the Centers for Disease Control and Prevention (CDC) Office on Smoking and Health (OSH), provides states guidance on conducting many aspects of the YTS such as questionnaire development, sample selection, administration of the survey and scanning the data.[1] The YTS Handbook contains a condensed description of the data processing tasks. Prior to July 2017, CDC performed data processing for the states. This assistance is no longer available. States that conduct the YTS must contract the data processing tasks or perform these tasks internally. This document provides specifications for performing these tasks.

The output of the data processing task includes an analytical dataset with derived variables and sampling weights, a codebook, frequency tables of analytical variables, and tables of analytic results. The inputs include the raw data file (output from scanning the survey booklets), the sample weights (output from PC Sample program), student population totals (from the states), along with other input data files.

[1] https://www.cdc.gov/tobacco/data_statistics/surveys/yts/pdfs/2011-yts-handbook.pdf

2. INPUT FILES

Each implementation of the survey (middle school or high school) requires 6 files. In the names of the file, "***st***" is the 2-character state abbreviation (e.g., nc, ms) and "***x***" is "m" for middle school data and "h" for high school data.

- File 1: ***stx***pblay.xls. This is the layout file created during the process of scanning in the data.
- File 2: YTS2017 ***ST X***PB RAWDATA.TXT. This is the raw data file created during the scanning of the data.
- File 3: ***stx***pbedt.xls. This defines the edits.
- File 4: ***stx***pbpop.xls. This is the population data used for the calibration totals. This file comes from the states.
- File 5: ***stx***pbprf.xls. This contains the definitions of the preferred (defined or calculated) variables.
- File 6: ***stx***pbwgt.xls. This contains the following data from the sample file: sampled schools, school enrollment, the school weights, the student weights; and the following data is provided from the state: an indicator for school response, the number classes, and for every class selected—class size and quantity of respondents.

3. PROCESSES

The State YTS data collection and analysis tool can be broken into six (6) different processes. These processes are listed below. Details for each process are provided in the associated section of this document.

- Section 4: Create a cleaned interim dataset in SAS format.
- Section 5: Check input files
- Section 6: Create sampling weights and creation of an analytical dataset in SAS format
- Section 7: Create several types of output for each preferred/analysis variable
- Section 8: Display statistics by sex, grade and race/ethnicity for each analysis variable
- Section 9: Create a codebook

4. CREATION OF A CLEANED INTERIM DATASET IN SAS FORMAT

This process produces a SAS dataset containing an interim version of the analysis data set. The interim version of the analysis dataset is identical to the final analysis dataset except it does not contain the sampling weights. The sampling weights are created in a separate program discussed in Section 6. This process also creates a text file containing checks of the interim data set. These checks are described in Section 4.15.

4.1 Structure of the layout file and raw data file

The Layout file, input file 1, is an Excel file containing the meta-data used to read the raw data. The name of the file is of the form ***STX***PBLAY.xls, where "***ST***" is a 2-character state abbreviation and "***X***" is "H" for high schools and "M" for middle schools. The layout file contains three variables names:

- NAME—the name of the variable on the original questionnaire, most of these variables are in character format;
- NEWNAME—the name of the analytical variable, most of these variables are in numeric format, and;
- ORIGQ—the numbering of the questions in the state questionnaire.

The variables POSITION and LENGTH indicate the position and length of the variable on the text file. The variable FREQTBL indicates if a "frequency table" will be created based on that variable. The variable QUESTION contains the label for the associated variable. The 13 variables ANS1—ANS13 contain the categories associated with the value of the variables read on the raw data file (i.e. ANS1 is associated with an answer of "A", ANS2 is associated with an answer of "B", valid values are the ones that are non-missing). This information is used when transforming the raw data into an analytical dataset.

The raw data file, input file 2, is a text file that contains the raw data. The name of this file is of the form YTS20***YR ST X***PB RAWDATA.TXT, where "***YR***" is the last 2 digits of the year, "***ST***" is the 2-character state abbreviation and "**X**" is "H" for high schools and "M" for middle schools. This file is the output to the processes that scan in the booklets.

4.2 Read in the raw data

The layout file provides the instructions on how to read in the raw data file. The values of the variables NAME and NEWNAME on the layout file are the variables on the analytic file.

The data on the raw data file are assigned the variables identified by the variable NAME on the layout file. The position of the variables on the raw data file is indicated by the value of POSITION on the layout file. The length of the variable on the raw data file is indicted by the value LENGTH on the layout file. The variable type, numeric or character, of the variable on the raw data file is indicated by the value TYPE on the layout file.

4.3 Add extra variables

Seven variables are added to the file. The code below indicates the variable name and the initialization of each variable.

```
Year=2017;          /* 4-digit year*/
Survey="YTS";
State="ST";         /*State abbreviation (e.g., NC)*/
SchoolType="YTS";
RegionName="State name"; /*(e.g., North Carolina)*/
RegionNumber=1;
keepObservation=1;
```

4.4 Make numeric variables

Most of the variables on the analytic file are in character format. The variable names came from the variable NAME on the layout file. For each character variable, we create a numeric variable named with the associated value of NEWNAME from the layout file. The following describes the process.

> For all character variables, if the value of NAME is A then value of NEWNAME is 1, if the value of NAME is B then value of NEWNAME is 2, …, if the value of NAME is M then value of NEWNAME is 13, and so forth.

4.5 Add labels

Labels are metadata that describe the variables. The layout file contains the labels for the variables identified by the variables NAME and NEWNAME. The labels contain the question associated with each variable. We add these labels to the analytic file.

4.6 Missing and out-of-range flags

For every variable in NEWNAME on the layout file, that was character in NAME on the layout file, we create two new variables; one is a missing flag and the other is an out-of-range flag. We first determine the valid values by observing the valid values of ANS1-ANS13 on the layout file; valid values are the ones that are non-missing. The variable name of the missing flag has the form MSXXX, where XXX is the variable name associated with NEWNAME on the layout file. The label of the variable MSXXX is “Missing Flag for XXX”, where XXX is the variable name associated with NEWNAME on the layout file. MSXXX = 1 if the value of MSXXX is not a valid value based on the nonmissing values of ANS1-ANS13 on the layout file; otherwise MSXXX = 0.

The variable name of the out-of-range flag has the form ORXXX, where XXX is the variable name associated with NEWNAME on the layout file. The label of the variable ORXXX is "Out of Range Flag for XXX", where XXX is the variable name associated with NEWNAME on the layout file. ORXXX = 1 if the value of XXX is not missing and has an invalid value based on the nonmissing values of ANS1-ANS13 on the layout file. Otherwise ORXXX = 0.

4.7 Add unused core variables

The derived variables and edit check variables are functions of the core variables. Not all the core variables are used in every iteration of the State YTS. We identify the core variables, which are required to define the derived variables and edit check variables, that were not asked in the State YTS. These variables are added to the analytical dataset and labeled "***UNUSED CORE VARIABLE***". For instance, in one State YTS the following core variables were not part of the questionnaire: CR9, CR10, CR11, CR16, CR19, CR21, CR22, CR27, CR34, CR38, CR39, CR63, CR70, CR71, CR72, CR90, CR91, CR94, CR112, CR113, CR135, CR136, CR137. Each of these variables is placed on the data file, all observations are initialized to missing and the label "***UNUSED CORE VARIABLE***" is applied.

4.8 Apply edit checks: create edit indicators and create edit fail indicators

The edit checks determine if the subject gave answers that are not logically consistent. File 3: ***stx***pbedt.xls contains the edit checks, where "***ST***" is a 2-character state abbreviation and "***X***" is "H" for high schools and "M" for middle schools. The variables from ***stx***pbedt.xls in the column EDITVAR are defined with the Boolean logic in the column CRITERIA. If the value of an edit check is true then (1) the second value in the expression in the column CRTIERIA is set to missing and (2) the edit failure indicator variable EFXXX is set to 1. Otherwise, if the value of an edit check is false, the edit failure indicator variable EFXXX is set to 0. For example, the edit check E1 is defined as 1 (true) if (CR1 IN (1, 2, 3, 4) and CR12 IN (7, 8, 9, 10, 11, 12, 13)). Note: true indicates inconsistent data. Edit E1 tests if a 9, 10, 11 or 12-year-old reported that they tried their first cigarette when they were 13 years-old or older. If E1 is true then CR12 gets assigned to missing and EFCR12 gets assigned to 1. If E1 is false then EFCR12 gets assigned to 0. The other edit checks are treated similarly.

4.9 Nonresponse flags

A nonresponse flag is created for all the variables with TYPE=”C” on the layout file. For example, for the variable CR1, a nonresponse flag variable is created with the name NRCR1. The label of the variable NRCR1 is “Nonresponse Flag for CR1.” NRCR1 is initialized to 1 if CR1 is missing (CR1=.) or the value of CR1 is out-of-range (ORCR1=1 as described in Section 4.6).

4.10 Create the derived variables

The definitions of the derived variables are found in File 5: ***stx***pbprf.xls, where “***st***” is a 2-character state abbreviation and “***x***” is “h” for high schools and “m” for middle schools. There are two steps to creating the derived variables. The derived variables are functions of the core variables. Not all the core variables are used in every iteration of the State YTS. The first step is to identify the core variables which are required to define the derived variables, that were not asked in the State YTS. These variables are added to the analytical dataset and labeled “***UNUSED CORE VARIABLE***”. The second step is to apply the logic described in File 5: stxpbprf.xls. For example, the derived variable ESOMKE is defined as follows: if CR7=1 then ESMOKE is initialized to 1; otherwise if CR7=2 then ESMOKE is initialized to 2. Note: CR7 refers to the question, “Have you ever tried cigarette smoking, even one or two puffs?” CR7=1 indicates “yes”; CR7=2 indicates “no.”

4.11 Susceptibility index

The Susceptibility Index is a concept developed by CDC using responses from the following five YTS questions. Note, these are the question numbers from the previous questionnaire (pre-2016/2017 update). States need to review their state questionnaire and update the coding to match up the correct variable names:

- CR7—Have you ever tried cigarette smoking, even one or two puffs?
- CR8—About how many cigarettes have you smoked in your entire life?
- CR9—Do you think that you will try a cigarette soon?
- CR10—Do you think you will smoke a cigarette in the next year?
- CR12—If one of your best friends were to offer you a cigarette, would you smoke it?

Depending on the responses, students are placed into one of four categories: non-susceptible never-smokers (NSNS), susceptible never-smokers (SNS), experimenters (Exp), and established smokers (EstS). The variable names in the list above come from the original questionnaire. Macro variables (&CR7SUSCEP.— &CR12SUSCEP.) are created to identify the question in the applicable State YTS associated with each of the questions that make up the susceptibility index. For example, if CR8 in the current State YTS is the question, “Have you tried cigarette smoking, even one or two puffs?" then we assign the macro variable &CR7SUSCP. to CR8. The following logic is applied.

```
LABEL SUSCEP = 'Susceptibility Index';
 if &CR7SUSCEP.=1 then do;
 if &CR12SUSCEP. IN (2,3,4,5,6,7) then SUSCEP=3; *experimenter;
 if &CR12SUSCEP.=8 then SUSCEP=4; *established;
 end;
 ELSE if &CR7SUSCEP.=2 then do;
 if &CR8SUSCEP.=4 and &CR9SUSCEP=3 and &CR10SUSCEP=4 then SUSCEP=1;
 *nonsusceptible, never smokers;
 else SUSCEP=2; *susceptible never smoker;
 if &CR8SUSCEP.=. And &CR9SUSCEP.=. And &CR10SUSCEP.=. Then SUSCEP=.;
 end;
```

4.12 Define RACE and number of multiple race/ethnicity selections

The following SAS code is used to recode race/ethnicity:
Race is defined as

```
      label RACE = "(1=WHITE) (2=BLACK OR AFRICAN AMERICAN) (3=HISPANIC OR
       LATINO) (4=OTHER)";
       if CR4 in (2,3,4,5) then RACE=3;
       else if CR5E=5 then RACE=1;
       else if CR5C=3 then RACE=2;
       else if CR5A=1 or CR5B=2 or CR5C=3 or CR5D=4 then RACE=4;
```

Number of multiple race selections is defined as:

```
      label Q5XN = "NUMBER OF MULTIPLE RACE SELECTIONS";
      Q5XN=sum(0,(CR5A ne .),(CR5B ne .),(CR5C ne .),(CR5D ne .),(CR5E ne
      .));
```

4.13 Dataset of deleted observations

An observation is deleted if the study is for middle schools and the student is not in grades 6 through 8 or the study is for high schools and the student is not in grades 9 through 12. A SAS dataset with the cleaned data is created called: ***STX***PB***YR***a.sas7bdat where "***ST***" is the 2-character state abbreviation, "***YR***" is the last 2 digits of the year and "**X**" is "H" for high schools and "M" for middle schools.

4.14 Output permanent interim cleaned dataset

A SAS dataset with the cleaned data is created called: ***STX***PB***YR***.sas7bdat where "***ST***" is the 2-character state abbreviation, "***YR***" is the last 2 digits of the year and "X" is "H" for high schools and "M" for middle schools.

4.15 Quality assurance and quality control

The QA/QC procedures are implemented throughout the code.

A text file is created called "2 ***YR ST*** Public ***X*** Schools Read_Edits Preferred Output.txt" where "***YR***" is the 4-digit year, "***ST***" is the full state name and "***X***" is "M" for middle school or "H" for high school. This file is useful for checking the creation of the cleaned data.

- QC check 4.1: Print each variable with both the character and numeric coding of the categories.
- QC check 4.2: Frequency table for each categorical variable with a cross of the numeric (NEWNAME) and character (NAME) values.
- QC check 4.3: Analysis variables not available for edit checks.
- QC check 4.4: Analysis variables not available for preferred analysis variables.
- QC check 4.5: Frequencies of additional variables after edits are applied.
- QC check 4.6: Edit variable analysis
- QC check 4.7: Frequency of the derived variables (aka. preferred variables) and the variables that were used to create them.
- QC check 4.8: Check the susceptibility index. Note that user input is required. In the program code reproduced below, the user must delete the macro variables that are not initialized to valid study variables.

```
proc freq data=a10 notitle compress;
 table SUSCEP*&CR7SUSCEP*&CR8SUSCEP*&CR9SUSCEP*&CR10SUSCEP/list missing;
run;
```

- QC check 4.9: List all unused preferred variables (all observations are missing).
- QC check 4.10: Nonresponse grid:
    - All analysis variables,
    - Only analysis variables with high nonresponse.

5. CHECK INPUT FILES

This process applies checks to the login or weights file (input file 6: ***stx***pbwgt.xls), the layout file (input file 1: ***stx***pblay.xls) and the raw data file (input file 2: YTS2017 ***ST X***PB RAWDATA.TXT). It creates a text file with the results of the checks of the form *3 "4-digit year" "state name" Public "High or Middle" Schools Input and Check Class Form OUTPUT.pdf*

5.1 QC login/weights file

- QC check 5.1.1: Print out schools with an indicator that the number of responding classes does not equal the number of classes selected.
- QC check 5.1.2: Check that the product of the school weight and the school interval is equal for all schools.

- QC check 5.1.3: For all schools, compare the expected number of respondents with the actual number of respondents.
- QC check 5.1.4: Check if the class size is smaller than the number of respondents in that class.

5.2 QC raw data file

- QC check 5.1.5: Check if there are duplicate records within a class
- QC check 5.1.6: Check that all observations have a valid class and school ID.

5.3 QC layout file

- QC check 5.7: Check that the layout file inputs all data to the last position.

5.4 QC consistency between files

- QC check 5.8: Check that the number of respondents in each class on the raw data file matches the numbers given on the login/weights file. Print out observations in classes that fail this check.
- QC check 5.9: Print the classes with zero respondents.
- QC check 5.10: Check that the number of respondents in each school on the raw data file matches the numbers given on the login/weights file.

6. CALCULATE SAMPLE WEIGHTS[2]

Weighting procedures may differ slightly for "atypical" samples which may involve additional stratification, oversampling or coordinated sampling with other school-based surveys.

6.1 General Description

A sample is drawn using a two-stage cluster sample design where schools are selected with probability proportional to school enrollment size and classes within schools are selected so that the overall probability of selection of each student is equal. Every eligible student has a chance of being selected.

A weight has been associated with each sample unit to reflect the likelihood of sampling each student and to reduce bias by compensating for differing patterns of nonresponse. The weight used for estimation is given by:

[2] The weighting procedures described in Section 6 are mostly copied verbatim from an internal CDC document titled "Weighting Procedures for the Youth Survey Revised 7/2006"

W = W1 * W2 * f1 * f2 *f3 *f4

W1 (school selection weight) = the inverse of the probability of selecting the school.

W2 (class selection weight) = the inverse of the probability of selecting a classroom within a selected school

f1 = a school-level nonresponse adjustment factor calculated by school size category (small, medium, large).

f2 = a nonresponse class adjustment factor calculated by school

f3 = a student-level nonresponse adjustment factor calculated by class

f4 = a post stratification adjustment factor calculated by gender and grade – OR – adjustment factor calculated by grade, gender, and race

6.2 Calculating W1 and W2

The base weights W1 and W2 come from the probabilities of selection for schools and then classes calculated during sample selection. It is important to understand the procedures for deriving the probabilities of selection. Calculating the overall sampling fraction is the first step.

Adjust class sizes and class participation

When the cleaned data set was created, respondents that are identified in an ineligible grade (grade 6-8 on a high school file or grade 9-12 on a middle school file) were removed. Input File 6 ***stx***pbwgt.xls is adjusted to reflect these ineligible sample members. For each ineligible respondent, the corresponding school and class has the class size reduced by one and the number of respondents reduced by one.

Overall Sampling Fraction

The overall sampling fraction is the probability of selecting a student. For example, a state wants to complete 1500 interviews with public middle school students. If the total public middle school enrollment is 150,000 students in 200 schools, then the probability of selecting any one student is theoretically 1500/150000 or 1/100 (.01). However, since we know we are not likely to get exactly 1500 interviews due to survey nonresponse at both the school and student levels the actual formula includes an adjustment to the total public middle school enrollment. As demonstrated below, the denominator is multiplied by the product of the expected student response rate and the expected school response rate. In our hypothetical example, suppose that the school response rate is expected to be 0.95 and the student response rate is thought to be 0.85 due to absences on the survey day or lack of parental permission to participate. Instead of 1/100, the overall sampling fraction is 1/(100*0.95*0.85)= 1/80.75 = 0.012384.

$$\text{Overall Sampling Fraction (f)} = \frac{\text{Total sampled students}}{\text{Total School Enrollment in Sampling Frame}}$$

or

$$\text{Overall Sampling Fraction (f)} = \frac{\text{Target sample size}}{\text{Student RR} * \text{ School RR} * \text{Total School Enrollment in Sampling Frame}}$$

Target Sample Size = Target Number of Students to Sample

Student RR = Expected Student Response Rate

School RR = Expected School Response Rate

This overall sampling fraction can be broken into component probabilities, the probability of selecting a school and the probability of selecting a classroom within the selected school.

First Stage Sampling Fraction (Probability of selecting a school)

The next step involves calculating the probabilities of selection for schools (first stage) and then classes (second stage). Schools are selected using systematic sampling with a random start and a method called PPS (Probabilities Proportional to Size). Probabilities of school selection are proportional to a measure of size (MOS) that is based on the enrollment for each school. Except for very large and very small schools, the measure of size is exactly equal to the enrollment in the target grades. Prior to sampling, schools on the frame are sorted in descending order of size. Depending on the number of schools that are to be sampled, some very large schools may be selected with certainty. As each “certainty” school is selected, it is removed from the frame of eligible schools.

Determining Certainty Schools

An initial sampling interval is calculated by dividing the sum of the enrollments of all the schools in the sample frame divided by the number of schools desired in the sample. YTS usually selects 50 schools for a sample of 1500 middle school students. If the total enrollment for all schools in the sample frame equals 150,000, then the sampling interval will be 3000 (150000/50=3000). Schools that have an enrollment greater than or equal to 3000 are treated as certainty schools and are removed from the sample frame. Certainty schools are always selected for the sample. Each time a school is selected with certainty and removed from the frame, the sampling interval is recomputed based on the enrollment of the schools remaining on the sampling frame and on the number of schools remaining to be selected.

$$\text{Revised sampling interval} = \frac{\text{Total school enrollment in revised frame}}{\text{Adjusted number of schools}}$$

Sampling of certainty schools continues until the enrollment of the largest school remaining on the frame is less than the revised sampling interval. At this point, sampling of certainty schools is complete. The Sampling Fraction for Certainty Schools is $f_{school} = 1$.

Determining Noncertainty Schools

If more schools are needed in the sample after the certainty schools have been selected, they are sampled from the schools remaining on the frame. The sampling procedure for these "noncertainty schools" includes adjustments to the measure of size for schools that have very small enrollments. This procedure ensures that each student has the same probability of selection for the sample and that this probability is equal to the overall sampling rate.

In our example, if there was one school with an enrollment of 3500, the new sampling interval would be 2990 (146500/49=2990) rather than 3000. There are no schools with an enrollment as large as 2990. Noncertainty schools are now selected using systematic sampling and the recalculated sampling interval. Probabilities for these schools are proportional to school enrollment. Special adjustments are made to MOS for very small schools, as described in the next section.

Adjusting for Schools with Very Small Enrollments

For noncertainty schools with very small enrollments, it is possible that the probability of selection based on enrollment is so small that the overall probability of selection for students in these schools may be less than the required overall rate. Specifically, this happens when the school enrollment is so small that, even if all students were selected with certainty, their probability of selection would not be equal to f, the overall probability. Therefore, an adjustment is made to the measure of size for small schools so that students from these schools will have the required overall probability of selection. Essentially, the probability of selection for the small schools is increased so that selecting students with certainty from these schools will match the overall probability of selection for students. This has the effect of slightly decreasing the probabilities of selection for the larger, noncertainty schools. The number of students sampled from the small schools is thus slightly larger than it would be ordinarily.

After the certainty schools have been removed from the sampling frame, the remaining schools are ordered according to decreasing enrollment. It then is necessary to determine a minimum measure of size, MINMOS for each school. When a school's enrollment falls below $MINMOS_i$, the actual enrollment is replaced by $MINMOS_{i.}$ The procedure that PC Sample uses for determining $MINMOS_i$ is relatively simple. For the first and largest school remaining on the frame $MINMOS_i = 0$ (remember the schools are sorted in descending order by size). Thereafter each succeeding school will have:

> $MINMOS_i$ = the overall sampling fraction f x the sum of the enrollment of all schools preceding $school_i$ / (number of schools remaining to be selected – (f x (number of schools in the adjusted frame – the number of schools preceding $school_i$ + 1)))

In a real-world example, a state had 737 schools with a total enrollment of 583,396 students on their sample frame. The state asked that schools with enrollments of twenty or less be eliminated and so the frame finally contained 723 schools with an enrollment of 583,313. The state wanted to select 60 schools to ensure a minimum of 2500 interviews from students. The overall sampling interval was 9,722 students, and no school exceeded this enrollment meaning that no certainty schools were selected from the frame. Even after eliminating schools with enrollments of twenty or less, there remained 14 schools with enrollments of less than fifty students. The overall sampling fraction equaled 0.0043. $MINMOS_i$ increases as the enrollment decreases. It reached a maximum of 52.10641 before the actual enrollment of the remaining fourteen schools fell below this number, the revised MOS for these fourteen schools became equal to the maximum MINMOS of 52.10641.

Selection of noncertainty schools is carried out using systematic sampling with a random start and an adjusted school sampling interval that uses a total enrollment based on the revised MOS_i.

$$\text{Adjusted school sampling interval} = \frac{\text{Total enrollment based on revised } MOS_i}{\text{Number of noncertainty schools required}}$$

The default selection procedure uses implicit stratification based on school enrollment. This procedure helps to ensure that schools of varying sizes are selected and helps to control the precision of estimates. In our example, the adjusted school sampling interval is 9,726 (583571/60) instead of 9,722.

Second Stage Sampling Fraction (Probability of selection within school)

Certainty Schools

Returning to our original example for simplicity, we see that all students in certainty schools have an overall probability of selection equal to the overall sampling rate (0.01 in our original example).

$$f_{\text{class}} = f$$

$$\text{Within School Sampling Interval} = \frac{1}{\text{Overall Sampling Fraction}} = 1/0.01 = 100$$

Noncertainty Schools

For noncertainty schools, the within school sampling probabilities are based on the adjusted school sampling interval and on the adjusted school probability of selection so that

$$\text{Adjusted school probability } (\text{f}_{school}) = \frac{\text{Adjusted school MOS}}{\text{Adjusted school sampling interval}}$$

and

$$\text{Within school interval} = \frac{\text{Adjusted school probability}}{\text{Overall sampling fraction}}$$

Assuming in our original example that no schools had an enrollment less than MINMOS_i, the school of 987 students would have a within school probability of selection of:

$$f_{\text{school}} = 987 / 2990 = 0.33$$

For a school with an enrollment of 600, the comparable calculation would be:

$$f_{\text{school}} = 0.600 / 2990 = 0.20$$

The within school intervals in each case are 33 (0.33/0.01) for the school with 987 students and 20 (0.20/0.01) for the school with 600 students.

The second stage probability for selecting classes within a school is calculated by dividing the overall sampling fraction by the school sampling fraction.

$f_{\text{class}} = 0.01 / 0.33 = 0.0303$ if the school has 987 students and

$f_{\text{class}} = 0.01 / 0.20 = 0.0500$ with an enrollment of 600 students.

The product of the first and second stage sampling fractions should equal the overall sampling rate. For the school with 987 students

$$f = f_{\text{school}} * f_{\text{class}} = 0.33 * 0.03 = 0.01$$

The corresponding product for the school with 600 students would be:

$$f = f_{\text{school}} * f_{\text{class}} = 0.20 * 0.05 = 0.01$$

This two-stage sampling procedure yields an overall sample with probability of selection for each student equal to the overall sampling fraction. Within each selected school the within school sampling interval is applied to a random start. For the school with 987 students the random start is, say, 22 (the random start must be within the sampling interval) and so the following classes are selected for the survey: 22, 55, 88, 121, 154, for example. It is expected that only the first the first one to four classes will be selected. However, the sampling interval is provided with the sample documentation so that more classes can be chosen if necessary.

W1 School Selection Weight

The school selection weight (W1) is the inverse of the school (first stage) sampling fraction. For the school with 987 enrollment the weight is

$$\frac{1}{\text{Probability of selecting a school}} = \frac{1}{0.33} = 3.030$$

For the school with 600 students, it is

$$\frac{1}{\text{Probability of selecting a school}} = \frac{1}{0.2} = 5.000$$

W2 Class Selection Weight

The class selection weight (W2) is the inverse of the class (second stage) sampling rate. The class selection weights for the two schools above are:

School with 987 students

$$\frac{1}{\text{Probability of selecting a class (student) within a school}} = \frac{1}{0.0303} = 33.000$$

School with 600 students

$$\frac{1}{\text{Probability of selecting a class (student) within a school}} = \frac{1}{0.05} = 20.000$$

Application of Non-Response Adjustments

An adjustment is made for school non-response, class non-response and student non-response. The purpose of the non-response adjustments is to refine the weights to adjust for bias from non-response.

6.3 f1 - School Non-Response Adjustment

Schools are placed into 3 categories – Large, Medium, and Small – using tertiles of enrollment. For each group, a school non-response adjustment factor is calculated as follows:

$$School\ nonresponse\ adjusment = \frac{\sum_{selected\ schools\ in\ tertile} school\ selection\ weight * school\ enrollment}{\sum_{participating\ schools\ in\ tertile} school\ selection\ weight * school\ enrollment}$$

6.4 f2 - Class Non-Response Adjustment

For situations where entire classes within a participating school do not respond, a class adjustment factor is computed for each school in the following manner:

$$Class\ adjusment\ factor = \frac{Number\ of\ classes\ selected}{Number\ of\ classes\ participating}$$

6.5 f3 - Student Non-Response Adjustment

Within each class, a student-level adjustment is made for students who are non-respondents. The adjustment is computed by doing the following:

$$Student\ adjustment\ factor = \frac{\sum_{eligible\ students} student\ weight}{\sum_{completed\ surveys} student\ weight}$$

Overall Non-Response Adjustment

Of the components of the overall non-response adjustment, the school adjustment factor will be the same for all units regardless of school or class. The class adjustment will be the same for all students in a school. The student adjustment factor will be the same for students within the same class. If all the selected classes participate, the class adjustment factor drops out of the calculation.

Overall non-response adjustment = (School adjustment factor) x (Class adjustment factor) x (Student adjustment factor).

6.6 Application of the Post-Stratification Adjustment

For the Post-Stratification, the sample data is adjusted to match the school population data using known frequencies. The post-stratification is done by gender and grade or by grade-race-gender. For students who have missing data on grade, race or gender the values are imputed in the following manner:

> For respondents missing gender, the respondent is a randomly assigned as a ‘male’ or ‘female’ value based on a uniform distribution of random numbers. For respondents missing grade, the respondent is assigned a grade value based on the average grade (rounded) of the student’s school. For respondents missing race, randomly assign the respondent a race value (White, Black, Hispanic, or Other)

based on the empirical distribution of the data. For example, if the respondents were 60% white, 60% of the respondents who are missing race will be randomly classified as white. Note that the values are only imputed for weighting purposes. The imputed values are not retained in the final data set.

$$Post-stratification\ adjustment = \frac{Population\ count_{grade-gender\ or\ grade-race\ category}}{Sum\ of\ adjusted\ weights_{grade-gender\ or\ grade-gender-race\ category}}$$

For YTS data sets, our intention is to post-stratify by grade-race-gender whenever possible.

6.7 Final Weight

Final Weight = (School Selection Weight) x (Class Selection Weight) x (Non-response Weight) x (Post-Stratification Adjustment).

6.8 Stratum and PSU Assignment

The assignment of Strata and PSUs is as follows

a. **Certainty Schools**

If the school was selected with certainty then the school forms its own stratum and the classes within the school are identified as the PSUs.

b. Non-Certainty Schools

The non-certainty strata are sorted from largest to smallest. The schools are then paired together to form each stratum. The schools within each Stratum are identified as the PSUs. In the case of an odd number of non-certainty schools, the last 3 smallest schools are placed into the same stratum.

6.9 QC checks applied during the weighting process

The QA/QC procedures are implemented throughout the code. A text file is created called "4 ***YR ST*** Public ***X*** Schools Weighting Output.txt" where "***YR***" is the 4-digit year, "***ST***" is the full state name and "***X***" is "M" for middle school or "H" for high school. This file is useful for checking the creation of the sampling weights.

- QC check 6.9.1: Print out the adjustment to class enrollment and class participation of each school and class combination that had the adjustment to class size and class participation applied (Section 6.2).

- QC check 6.9.2: Print out all schools where PARTICIP=”.”. These are the schools that were selected to participate in the study but stopped operating as a school before data collection occurred.
- QC check 6.9.3: Print out all nonresponding schools.
- QC check 6.9.4: Check that all school and class combinations do not have more respondents than eligible students.
- QC check 6.9.5: Print out assignment to analysis stratum for each school.
- QC check 6.9.6: Print out school, student and overall response rate.
- QC check 6.9.7: Print out school by weighting class with response indicator enrollment and sum of enrollment by weighting class.
- QC check 6.9.8: Within weighting class, for each class, print out the nonresponse adjustment for school and student. Also, print out the total weight with the nonresponse adjustments.
- QC check 6.9.9: Examine the distributions of the nonresponse adjustments. Print out any large adjustments.
- QC check 6.9.10: Check that the sum of the adjusted weights and the total population are close, within 5%.
- QC check 6.9.11: Print out all imputed calibration variables.
- QC check 6.9.12: Compare the unweighted distribution of the calibration variables before and after imputation.
- QC check 6.9.13: Compare the weighted distribution of the calibration variables before and after imputation.
- QC check 6.9.14: Compare the weighted distribution of the calibration variables after calibration with the population totals. Determine that they are identical.
- QC check 6.9.15: Print the post-stratification adjustment factor for each post-stratification cell.
- QC check 6.9.16: Print cases where the final weight is less than 1.
- QC check 6.9.17: Print quantity of respondents by stratum and PSU.
- QC check 6.9.18: Print strata with less than 1 PSU.

- QC check 6.9.19: Check that all observations have a nonmissing stratum, PSU and final weight.

7. FOR EACH PREFERRED/ANALYSIS VARIABLE, CREATE SEVERAL TYPES OF OUTPUT

In this process, 3 types of tables are created and output to a text file called "5 ***YR ST*** Public ***X*** Schools Preferred Response (Gender, Grade and Race) Output.txt" where "***YR***" is the 4-digit year, "***ST***" is the full state name and "**X**" is "M" for middle school or "H" for high school.

7.1 For each preferred analysis variable, display statistics by sex, grade and race/ethnicity

There are approximately 44 preferred analysis variables. A table could be generated for each analysis variable, displaying for each level of that variable and in total, the percentage, margin of error and sample size and total by sex, grade and race/ethnicity. The following is an example of a table for the preferred analysis variable ESMOKE.

```
      Table 1.   Weighted Frequency Report for State
      Region 1 Public High Schools -- Based on Standard Survey YTS2017
      z Distribution Method -- Weighted Percents and Unweighted Sample Sizes for Each Column
      Reported by Gender, Grade and Race
      Analysis Variable ESMOKE -- Percent ever smoked cigarettes
                                       Gender/Grade/Race

                                                                                     Black or Hispanic
                                                                                     African     or
              Total    Female    Male     9th      10th      11th      12th     White    American  Latino    Other

Yes

   %             28.1     27.0     28.8     16.0     23.0     30.7     42.7     27.4     23.0     36.4     31.9

   95% CI         6.1      7.9      5.2      2.6      3.6      8.7      6.5      5.5      8.0      8.2     17.8

   n              574      264      300       89      253      112      119      371       55       80       66

No

   %             71.9     73.0     71.2     84.0     77.0     69.3     57.3     72.6     77.0     63.6     68.1

   95% CI         6.1      7.9      5.2      2.6      3.6      8.7      6.5      5.5      8.0      8.2     17.8

   n            1,620      816      796      429      813      222      155    1,152      181      142      133

Total

   %            100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0

   95% CI         0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0

   n            2,194    1,080    1,096      518    1,066      334      274    1,523      236      222      199
```

7.2 For each preferred analysis variable, create a crosstab of the preferred variable with the variables used to define that variable (preferred codebook)

For each of the approximately 44 preferred analysis variables we can create a table that contains a crosstab of the preferred analysis variable with the variables that are used to define the preferred analysis variable. The following is an example of the table for the analysis variable AGEGRP. We used a small font so the tables fit into the document.

```
Analysis
Variable                Question Code and Label                 Frequency      Percent     95% CI

AGEGRP     (Table 14) Percent of students under age 18              2,240         81.43%  ( 73.9- 88.9)
           IF (CR1 IN (1,2,3,4,5,6,7,8,9)) THEN AGEGRP = 1;         2,020         90.18%
           ELSE IF (CR1 IN (10,11)) THEN AGEGRP = 2;                  220          9.82%

                   AGEGRP    CR1    COUNT    PERCENT
                      .       .       12      0.5329
                      1       1        1      0.0444
                      1       5        2      0.0888
                      1       6      184      8.1705
                      1       7      747     33.1705
                      1       8      776     34.4583
                      1       9      310     13.7655
                      2      10      197      8.7478
                      2      11       23      1.0213
```

7.3 For each preferred analysis variable, display the weighted percents and unweighted sample sizes

For each of the approximately 44 preferred analysis variables we can create a table that contains the weighted percents and unweighted sample sizes. The following is an example of part of the table.

```
                  Weighted Preferred Summary Report for State
                     Region 1 Public High Schools -- Based on Standard Survey YTS2017
z Distribution Method -- Weighted Percents and Unweighted Sample Sizes for Each Preferred
Analysis Variable

                               VARNAME           n      %         95% CI

                               CORE134A        172    83.76    ( 74.4- 93.1)
                               AGEGRP        2,240    81.43    ( 73.9- 88.9)
                               CORE124A        180    78.38    ( 72.8- 83.9)
                               CORE134B        523    55.37    ( 46.6- 64.1)
```

8. FOR EACH ANALYSIS VARIABLE, DISPLAY STATISTICS BY SEX, GRADE AND RACE/ETHNICITY

Tables are created and outputted to a text file called “6 ***YR ST*** Public ***X*** Schools Frequency Tables (Gender, Grade and Race) Output.txt” where “***YR***” is the 4-digit year, “***ST***” is the full state name and “***X***” is “M” for middle school or “H” for high school.

8.1 For each analysis variable, display statistics by sex, grade and race/ethnicity

There are approximately 66 analysis variables; however, the exact quantity of analysis variables depends on the number of questions included in the state questionnaire. A table can be created for each analysis variable, displaying for each level of that variable and in total, the percentage, margin of error and sample size and total by sex, grade and race/ethnicity. The following is an example of a table for the analysis variable for question 4— "Do you currently get free or reduced-price lunch at school?"

```
Table 4.   Weighted Frequency Report for State
Region 1 Public High Schools -- Based on Standard Survey YTS2017
z Distribution Method -- Weighted Percents and Unweighted Sample Sizes for Each Column
Reported by Gender, Grade and Race
Analysis Variable MNR4 -- (Q4) Do you currently get free or reduced-price lunch at school?

                                                                        Gender/Grade/Race

                                                                                                    Black or Hispanic
                                                                                                    African     or
                               Total    Female    Male     9th      10th     11th     12th    White  American  Latino   Other

 Yes

    %                           37.1     38.0     35.6     36.3     34.7     33.0     43.8     21.2     84.7     67.7     69.9

    95% CI                       8.0      9.0      7.9      6.8     11.8      7.8     10.5      3.9      8.0      7.4     18.1

    n                            837      411      415      198      387      120      128      322      201      154      149

 No

    %                           62.9     62.0     64.4     63.7     65.3     67.0     56.2     78.8     15.3     32.3     30.1

    95% CI                       8.0      9.0      7.9      6.8     11.8      7.8     10.5      3.9      8.0      7.4     18.1

    n                          1,381      676      697      327      688      214      152    1,211       37       73       53

 TOTAL

    %                          100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0    100.0

    95% CI                       0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0      0.0

    n                          2,218    1,087    1,112      525    1,075      334      280    1,533      238      227      202
```

9. CREATE CODEBOOK

A codebook is created and outputted to a text file called "7 ***YR ST*** Public ***X*** Schools Codebook Output.pdf" where "***YR***" is the 4-digit year, "***ST***" is the full state name and "***X***" is "M" for middle school or "H" for high school.

For each analysis variable the following is displayed

- the original question name,
- the analysis variable name,
- the question label,

For each category of the variable the following is displayed

- the levels of the variables
- the labels of the levels

- unweighted frequency
- weighted percentage