## Appendix N. Detailed Sampling and Weighting Plan

**SAMPLING AND WEIGHTING PLAN**
**NATIONAL YOUTH TOBACCO SURVEY**

This study will employ a repeat cross-sectional design to develop national estimates of tobacco use behaviors and exposure to pro- and anti-tobacco influences among students enrolled in grades 6-12. The study represents the continuation of the NYTS cycles that took place in 1999, 2000, 2002, 2004 and 2006. The NYTS came before OMB for the first time in 2003 after management of the survey was passed from Legacy to CDC.

As presented in this supporting justification, every effort has been made to maintain the methodology established in prior cycles of the NYTS to permit comparability across cycles. The primary objectives of the NYTS are to develop estimates of tobacco use behaviors and exposure to pro- and anti-tobacco influences among students enrolled in middle school and high school grades; to identify differences related to demographic characteristics (age, grade, gender, and race/ethnicity); and to determine whether there are time trends in tobacco use behaviors and exposure to influences that promote or discourage tobacco use. Such information is required to support CDC's responsibilities in providing technical assistance in the planning, monitoring, and evaluation of national, state, and local tobacco prevention and control programs.

We propose to replicate key aspects of the sampling design followed in the 2006 NYTS. Refinements typically occur in response to the changing demographics of the in-school population and to meet CDC's policy needs. For example, current trends of increasing percentages of minority students likely will lead to more efficient sampling of minority students and to smaller overall sample sizes. The planned design increases the representation of black and Hispanic students by oversampling schools and areas with high concentrations of these minority groups.

## 1. Estimation and Justification of Sample Size

### *1.1 Overview*

The NYTS is designed to produce the key estimates accurate to within ± 5% at a 95% precision level. Estimates by grade, gender, and grade by gender meet this standard. The same standard is used for the estimates for racial/ethnic groups by school level.

### *1.2 Expected Precision*

Confidence intervals vary depending upon whether an estimate represents the full population or a subset, such as a particular grade, gender, or racial/ethnic group. Within a grouping, they also vary depending on the level of the estimate and the design effect associated with the measure. Based on the prior NYTS studies, as well as on precision requirements that have driven the sampling design, we can expect the following subgroup estimates to be within ± 5% at 95% precision level:

- Estimates by grade, gender, and grade by gender

- Minority group estimates by school level (middle school versus high school) for black and Hispanic students

These precision estimates are based on the following projected numbers of participating students: 2,840 students per grade (with a 50% gender distribution); 1,583 black students per school level; and 1,583 Hispanic students per school level. For conservative design effect scenarios (design effects of 2.0 or less), estimates based on these subgroup sample sizes will be within +/- 5 percentage points at the 95% confidence level.

Exhibit N-1 summarizes the sample sizes planned for each school type. These sample sizes will generate approximately 2,840 participating students per grade. In this exhibit, the first two columns show the number of school selections and student selections per school while the last column shows the expected number of participating students. The numbers of large, medium and small schools were carefully calibrated to generate the required numbers of students overall, by school level and by grade.

**Exhibit N-1 Sample Size Projections for Participating Students on the 2009 NYTS**

| School Type | Number of Schools Selected | Number of Students Selected per School | Expected Number of Participating Students Across Schools |
|---|---|---|---|
| Large, High Minority High Schools | 60 | 200 | 8,670 |
| Large, Non-High Minority High Schools | 20 | 100 | 1,445 |
| Medium High Schools | 10 | 100 | 723 |
| Small High Schools | 10 | All available (assume 72) | 520 |
| Large, High Minority Middle Schools | 60 | 150 | 6,502 |
| Large, Non-High Minority Middle Schools | 20 | 75 | 1,085 |
| Medium Middle Schools | 10 | 75 | 542 |
| Small Middle Schools | 10 | All available (assume 54) | 390 |
| **TOTAL** | **200** | | **19,877** |

As summarized in Exhibit N-2, the sample design will yield 1,763 and 1,583 black students for high school and middle school, respectively. The exhibit also shows that somewhat higher numbers are expected for Hispanic students, 1,991 for high schools and 1,753 for middle schools.

**Exhibit N-2 Expected Minority Yields**

Part I: High Schools

| School Strata | Number of Schools | Expected Number of Black Students | Expected Number of Hispanic Students |
|---|---|---|---|
| Large HS, High-Minority | 60 | 1387.2 | 1560.6 |
| Large HS, Non-High Minority | 20 | 202.3 | 231.2 |
| Medium HS | 10 | 101.15 | 115.6 |
| Small HS | 10 | 72.828 | 83.232 |
| TOTAL | 100 | 1763.478 | 1990.632 |

**Part II: Middle** Schools

| School Strata | Number of Schools | Expected Number of Black Students | Expected Number of Hispanic Students |
|---|---|---|---|
| Large MS, High-Minority | 60 | 1300.5 | 1430.55 |
| Large MS, Non-High Minority | 20 | 151.725 | 173.4 |
| Medium MS | 10 | 75.8625 | 86.7 |
| Small MS | 10 | 54.621 | 62.424 |
| TOTAL | 100 | 1582.709 | 1753.074 |

*1.3 School and Student Non-response*

Across the five cycles of the NYTS, the school participation has averaged 89%, with a low of 83%. Student participation has averaged 90% with a low of 88%. To be conservative, we have assumed slightly lower values in developing the sample design for the 2009/2011 NYTS: 85% for schools and 85% for students.

## 2. Sampling Design

*2.1 Overview*

The sampling universe for the NYTS will consist of all public, Catholic and other private school students in grades 6 through 12 in the 50 States and the District of Columbia. The sampling frame for schools will be constructed using files obtained from Quality Education Data, Inc. (QED). QED data encompasses both private and public schools and includes the latest data from the Common Core of Data from the National Center for Education Statistics.

The NYTS design effectively oversamples black and Hispanic students by increasing the sampling intensity in those schools with high concentrations of these minority groups. More specifically, the oversampling is achieved in three different ways: 1) by allocating the primary

(PSU) sample disproportionately to those strata that have higher concentrations in these minority groups; 2) by using a weighted measure of size in selecting PSUs with probabilities proportional to size (PPS) that assigns greater probabilities to high-concentration schools and areas; and 3) by selecting more classes in those schools that have high concentrations in these minority groups.

The sample will be a stratified, three-stage cluster sample stratified by racial/ethnic status and urban versus rural. PSUs are classified as "urban" if they are in one of the 54 largest MSAs in the U.S.; otherwise, they were classified as "rural". Additional, implicit stratification will be imposed by geography by sorting the PSU frame by state and by 5-digit Zip Code (within state). Within each stratum, a primary sampling unit (PSU), defined as a county or a group of counties, will be chosen without replacement at the first stage. In subsequent sampling stages, a probabilistic selection of schools and students will be made from the sample PSUs. Exhibit N-3 presents a summary of the sampling design features.

**Exhibit N-3 Key Sampling Design features**

| Sampling Stage | Sampling Units | Sample Size (Approximate) | Stratification | Measure of Size |
|---|---|---|---|---|
| 1 | PSUs: Counties or groups of counties | 80 | Urban vs. non-urban (2 strata); Minority concentration (8 strata) | Aggregate school size in target grades |
| 2 | Schools | 200 school selections: 160 large schools (2 per PSU), 20 medium schools and 20 small schools | Small, Medium and Large; High-school vs. middle-school | Weighted enrollment (increased for black, Hispanic groups) |
| 3 | Classes/ students | 1 or 2 classes per grade (2 per grade in large high-minority schools) 27,510 selected students | | |

We plan to utilize 80 PSUs, randomly selecting two large schools (one high school and one middle school) from each PSU. In addition, we will randomly select two “small” schools (one high school and one middle school) and two “medium” schools (one high school and one middle school) in each of a randomly selected subset of 10 PSUs. Including the small and medium schools, a total of 200 schools will be selected, with the expectation that approximately 85% or 170 of these schools will participate in the survey. Because some schools may not have a complete set of grades for the school level (middle school or high school), the second-stage

units (SSUs), are sometimes combinations of schools.

Small schools will be defined first as those schools with 25 or fewer students in one of the eligible grades. Of schools that are not small, medium schools will be defined as having fewer than 50 students in one of the eligible grades for the level. The remaining schools will be considered large. We plan to oversample black and Hispanic students by using a modified measure of size that will increase the probability of selection of schools with disproportionately high minority student enrollments. In addition, more classes will be selected in large schools with the highest enrollment of minority students.

In large schools, we will select an average of 1.75 classes per grade by selecting 2 classes per grade in 75% of the selected large schools, and one class per grade in the remaining schools. The double class sampling will take place in the 60 sampled schools with the highest concentrations of Hispanic and black students. Specifically, the sample will include 60 high-minority, large high-schools and 20 non-high minority, large high-schools. "Non-high minority" does not imply that the concentrations of minorities in these 20 schools is low by an absolute standard; rather, it only means they are lower than the top 60, regarded as high-minority. Similarly, it will contain 60 high-minority, large middle-schools and 20 non-high minority, large middle-schools.

### *2.2 Measure of Size*

The sampling approach will utilize Probability Proportional to Size (PPS) sampling methods to achieve over-sampling of black and Hispanic students. In PPS sampling, when the measure of size is defined as the count of final-stage sampling units, and a fixed number of units are selected in the final stage, the result is an equal probability of selection for all members of the universe. For the NYTS, we approximate these conditions, and thus obtain a roughly-self weighting sample. This section describes the type of measure of size to be employed for selecting PSUs and schools to over-sample black and Hispanic students.

One way of accomplishing over-sampling is to use a modified measure of population size during the PPS sample selection steps. A function of the form $r_hH + r_bB + r_oO$ is sought where the r's are the weighting factors for the Hispanic, black, and other populations (H, B, and O, respectively). Note that the "other" group includes white students and students of other race/ethnicity who are selected at lower rates than black and Hispanic students. This function will increase the chances of schools with relatively large minority enrollments entering the sample.

As in previous NYTS cycles, the coefficients for the measure of size weighting function will be developed using simulation studies that ensure that target sample sizes are met for the two racial/ethnic groups of analytic interest (black and Hispanic students). In 1990, Macro conducted a series of simulation studies that investigated the relationship of various weighting functions to the resulting numbers and percentages of minority students in the obtained samples.[1] Starting with coefficients used in previous cycles, these simulations are performed in iterative steps that 1) adjust for racial/ethnic composition, and 2) equalize the expected sample size by

1 Errecart, M.T., Issues in Sampling African-Americans and Hispanics in School-Based Surveys. Centers for Disease Control, October 5, 1990.

grade. Prior to full implementation, sampling design parameters are fine-tuned through simulated sample draws. The simulations involve the repeated selection of independent samples of schools (e.g., 1,000 samples) in an iterative application of the entire sampling process up to the selection of schools.

The effectiveness of a weighted measure of size in achieving oversampling is dependent upon the distributions of black and Hispanic students in schools. For example, if U.S. schools had identical percentages of minorities in every school, then the sample of students from any sample of schools would mirror the national percentages and use of a weighted measure of size would fail to oversample blacks and Hispanics. We know this is not the case, however, as the distribution of high school students with respect to race and ethnicity follows that of the general population, and here we find a great deal of clustering by race and ethnicity. Application of a weighted measure of size in prior cycles of NYTS has been effective in oversampling blacks and Hispanics.

The measure of size will be used also to compute stratum sizes and PSU sizes. Assigning an aggregate measure of size to strata, the sample allocation will oversample strata that have higher minority student concentrations, and further increase the chances for high-minority PSUs getting into the sample.

### *2.3 Definition of Primary Sampling Units (PSUs)*

In defining PSUs, several issues are considered:

a. Each PSU should be large enough to contain the requisite numbers of schools and students by grade, yet not so large as to be selected with near-certainty.

b. Each PSU should be compact geographically so that field staff can go from school to school easily.

c. There should be recent data available to characterize the PSUs.

d. PSU definitions should be consistent with secondary sampling unit (school) definitions.

Generally, counties will be equivalent to PSUs with two exceptions: (1) low population counties are combined to provide sufficient numbers of schools and students, and (2) counties that are very large may be split to avoid becoming certainty or near-certainty PSUs. County population figures will be aggregated from school enrollment data for the grades of interest. Enrollment data are being obtained from the most recent Common Core of Data from the National Center for Education Statistics and the current school and school district data files of Quality Education Data, Inc. (QED), which are updated continuously.

The PSU frame for the 2009 and 2011 NYTS will be formed directly from counties using methods developed by Macro. The methods employ both student counts and geographic data to ensure that the PSUs have the correct number of schools and students, and that the PSUs are compact geographically.

*2.4 First-stage Sampling: Stratification and Selection of PSUs*

### 2.4.1 Definition of Strata

The PSUs will be organized into 16 strata, based on urban/rural location (as defined above) and minority enrollment. The approach involves the computation of optimum stratum boundaries using the cumulative square root of "f" method developed by Dalenius and Hodges. The boundaries or cutoffs change as the frequency distribution ("f") for the racial groupings change from one survey cycle to the next. These rules are summarized below.

If the percentage of Hispanic students in the PSU exceeds the percentage of black students, then the PSU is classified as Hispanic. Otherwise it is classified as black. If the PSU is within one of the 54 largest MSA in the U.S. it is classified as 'urban', otherwise it is classified as 'rural'. Hispanic urban and Hispanic rural PSUs were classified into four density groupings depending upon the percentages of Hispanics in the PSU. Black urban and black rural PSUs were also classified into four groupings depending upon the percentages of blacks in the PSU.

### 2.4.2 Allocation of the PSU sample

We will design and select a sample of 80 PSUs. In order to stay as close as possible to maximum sample efficiency in terms of precision, the initial allocation of PSUs will be made proportional to student enrollment. Then, we will make adjustments to the initial allocation to meet minority targets, and evaluate these adjustments using sample simulations. This entire process is similar to that used in prior cycles of NYTS. Response rates from prior cycles will be used to inform the yield computations in the simulations.

### 2.4.3 Selection of PSUs

- Within each first-stage stratum, the PSUs will be sorted by five-digit zip code to attain a form of implicit geographic stratification. Implicit stratification, coupled with the probability proportional to size (PPS) sampling method described below, will ensure geographic sample representation. With PPS sampling, the selection probability for each PSU is proportional to the PSU's measure of size. The following systematic sampling procedures will be applied to the stratified frame to select a PPS sample of PSUs.

- Select 80 PSUs with a systematic random sampling method within each stratum. The method applies within each stratum a sampling interval computed as the sum of the measure of size for the PSUs in the stratum divided by the number of PSUs to be selected in the stratum.

- Subsample at random 10 of the 80 sampled PSUs for small school sampling and 10 of the 80 sampled PSUs for medium school sampling.

*2.5 Second Stage - Selection of Schools*

Schools will be stratified by school level—middle schools and high schools—and by size. We will define three school strata-- small, medium and large. Small schools will be defined first as those schools with 25 or fewer students in one of the eligible grades. Of schools that are not small, medium schools will be defined as having fewer than 50 students in one of the eligible grades for the level. The remaining schools will be considered large.

In addition, for each school level, large schools will be classified as high-minority (top 75% of the 60 large schools selected) or low-minority. High-minority schools initially will be identified as those containing more than a specified percentage of black or Hispanic students; e.g., the upper quartiles of the distribution of black and Hispanic students may be used as thresholds for high-minority schools. Then, the threshold will be fine-tuned so that approximately 75% of the large schools sampled are classified as high-minority schools.

For large schools, one high school and one middle school will be selected with PPS systematic sampling within a PSU. The schools will be selected into the sample with probability proportional to the weighted measure of size.

Small and medium schools will be sampled independently from large schools; they will be selected by drawing two mutually exclusive subsamples of PSUs. Ten schools will be selected within each subsample. In one subsample, one medium high school and one medium middle school will be drawn from each of the 10 PSUs. In the other subsample, one small high school and one small middle school will be drawn from each of the 10 PSUs. Medium and small schools will be selected in each sub-sampled PSU with probability proportional to size, using the weighted measure of size.

*2.6 Third Stage - Selection of Grades and Classes*

Except for cluster schools, classes are selected from all eligible grades in each school. In school clusters, grade samples are selected independently with one component school being selected for each grade.

The method of selecting students will vary from school to school, depending upon the organization of that school and whether a cluster of schools is involved. The key element of the school sampling strategy is to identify a structure that partitions the students into mutually exclusive, collectively exhaustive groupings that are of approximately equal sizes and that are accessible. Beyond that basic requirement, we will do the partitioning to result in groups in which both genders and students of all ability levels are represented. In selecting classes, we will generally give preference to selecting from mandatory courses such as English. Another option is to select from all classes that meet during a particular time of day such as all second or third period classes.

We will not use special procedures to sample for minorities at the school building level for two reasons:

- Schools do not maintain student rosters that identify students by racial/ethnic affiliation.
- We feel this would be viewed by many schools as an offensive practice.

We plan to select one or two classes per grade from each participating school. Two classes per grade are selected in schools classified as “high minority.” In the case of school clusters, we will conduct our sampling on a grade by grade basis. At each grade we will determine the identity of all schools in the cluster with students in that grade. If each school has enough students in the grade, then we will pick randomly one of the schools with probability proportional to grade enrollment and then select all of the classes from that school. If one of the schools does not have enough students, then its students will be combined with a class of another school in the cluster. If that class is picked, then students from both schools are selected to participate.

A "class" will be defined by our sampling team so that it meets size and composition requirements before the sampling is done. For example, two small classes may be combined and treated as one for sampling purposes. Or, boys and girls physical education classes may be combined. This approach is an efficient method of data collection in schools and also has the advantage of using the classroom teacher to distribute consent forms; hence, it tends to yield higher student participation rates. The disadvantage of this approach is that the sampling design tends to be less efficient because students within a class section tend to be more homogeneous than the entire student population within a school. The effect of this inefficiency has been accounted for in our estimates of the design effect of the study.

### *2.7 Replacement of Schools/School Systems*

We will not replace refusing school districts, schools, classes, or students. We have allowed for school and student non-response by inflating the sample sizes to account for non-response. With this approach, all schools can be contacted in a coordinated recruitment effort, which is not possible for methods that allow for replacing schools.

### *2.8 Selection of Students*

All eligible students in a selected class will be invited to take the survey.

## 3. Weighting

The final data set will be weighted to reflect the initial probabilities of selection and non-response patterns, to mitigate large variations in sampling weights, and to post-stratify the data to known control characteristics.

### *3.1 Initial Weighting*

The basic weights will be computed as the reciprocal of the probability of selection of that case. If $k_i$ is the number of PSUs to be selected from stratum i, $N_i$ is the size of stratum i and $N_{ij}$ is the size of PSU j in stratum i (in all cases size refers to our proposed measure of size), then the probability of selection of PSU j is $k_i N_{ij}/N_i$.

Assuming that one school is to be selected in stratum i, $N_{ijr}$ is the size of school r in PSU j in stratum i, then the conditional probability of selection of the school given the selection of the PSU is $N_{ijr}/N_{ij}$. If $C_{ijr}$ is the number of classes in school ijr and m is the number of classes to be selected then the conditional probability of selection of a class is $m/C_{ijr}$. Since all students are selected, the conditional probability of selection of a student given the selection of the class is unity.

The overall probability of selection of a student in stratum-i is the product of the three conditional probabilities of selection:

$$k_i N_{ij}/N_i$$
$$N_{ijr}/N_{ij}$$
$$m/C_{ijr}$$

This product, the student probability of selection, simplifies to a factor proportional to $k_i N_{ijr}/N_i C_{ijr}$

Therefore, the probabilities of selection will be the same for all students in a given school.

### *3.2 Non-response Adjustments*

Several adjustments are planned to account for student and school non-response patterns. First, a student non-response adjustment at the school level will be made by multiplying the student weight by the ratio of the total number of students in eligible grades to the total number of responses from students in the school.

An overall adjustment for student and school non-response will be made by stratum. This will adjust the sum of the student weights to the weighted measure of size of that stratum.

### *3.3 Weight Trimming*

Extreme variation in sampling weights can cause inflated sampling variances, and offset the precision gained from a well-designed sampling plan. This variation can occur, for example, if the number of respondents in a particular school is very low, causing a large non-response adjustment factor to be computed. One strategy to compensate for this is to trim extreme weights and distribute the trimmed weight among the untrimmed weights.

The trimming procedure will be iterative. In each iteration an optimal weight, $W_o$ is calculated from the sum of the squared weights in the sample. Then, each weight $W_i$ is marked and trimmed if it exceeds that optimal weight. The trimmed weight is summed within grade and spread out proportionally over the unmarked cases in the grade. This process is repeated until little or no weight is being trimmed. Weight trimming is done within stratum.

Typically, 2 to 3 percent of the total sample weight is trimmed and redistributed under the

weight trimming procedure.

### *3.4 Post-stratification to National Estimates*

National estimates of the student distribution by race/ethnicity, gender, and school level (middle vs. high school), by stratum, will be obtained from the QED database and NCES data (CCD for public schools and the Private School Survey for non-public schools). The QED files have five racial/ethnic categories: white; black; Hispanic; Asian and Pacific Islander; and Native American. As the survey data distinguish these categories, they will be used in post-stratification. (Note that these categories can be different from those considered in the design.)

Given a national estimate of $R_a$ and a weighted population estimate of $P_a$ for race category a in some grade-gender combination, the simple post-stratification factor would be the ratio of $R_a$ to $P_a$ for each gender, race and grade. The simple adjustment will be modified so that the percentage of respondents in the missing category remains constant.

If possible, this post-stratification will be done by sampling stratum. In other words, post-stratum cells will be constructed within each of the primary strata defined by minority concentration and urban status. We will collapse post-stratum cells that may be considered too small, i.e., those with too few participating students.

## ***4. Survey Estimators and Variance Estimators***

This section describes the computation of weighted survey estimators and the appropriate measures of sampling variability, or sampling error. These measures, such as variances and standard errors, will support the construction of confidence intervals and other statistical inference such as statistical testing.

If $w_i$ is the weight of case i (the inverse of the probability of selection adjusted for nonresponse and poststratification adjustments) and $x_i$ is a characteristic of case i (e.g., $x_i$=1 if student i smokes, but is zero otherwise), then the mean of characteristic x will be $(\Sigma\ w_i x_i)/(\Sigma\ w_i)$. A population total would be computed similarly as $(\Sigma\ w_i x_i)$. Weighted population estimates will be computed with the Statistical Analysis System (SAS) and SUDAAN software.

Sampling variances that account for the complex sampling design will be estimated using the method of general linearized estimators[2] as implemented in the SUDAAN[3] or SAS Version 8.2 survey procedures. These software packages permit estimation of sampling variances for multistage stratified sampling designs, accounting for unequal weighting, and for sample clustering and stratification. These software packages require the specification of sampling stages and sampling parameters (strata, PSU).

---

2 Skinner CJ, Holt D, and Smith TMF, Analysis of Complex Surveys, John Wiley & Sons, New York, 1989, pp. 50.

3 Shah BV, Barnwell GG, Bieler GS. SUDAAN: software for the statistical analysis of correlated data, release 7.5, 1997 [user's manual]. Research Triangle Park, NC: Research Triangle Institute; 1997.