**YHS 2017 SURVEY METHODOLOGY**

**I. Sample Design and Selection**

**A. High Schools**

In 2017, as in previous years, the population under study for this survey was all ninth through twelfth grade students attending a public high school in Massachusetts. The sample of high schools was selected by the CDC using sampling procedures consistent with what they use for all states. This is a probability proportionate to size selection of high schools with classes selected within high schools in such a manner that it leads to a self-weighting sample design for students throughout the state. In the case of Massachusetts, the probability of a student being selected for the survey was 0.029234. A total of 79 high schools were selected. DESE then provided contact information for the principal of each of these 79 schools. Since participation in these surveys is not mandatory, each of these 79 schools needed to be contacted to enlist cooperation. These efforts will be discussed later.

The sample design next called for obtaining a list of all required English or Social Studies classes from each selected high school. Using this list and directions supplied by the CDC, a systematic selection of classes from within each school was performed for each of the YRBS and the YHS. The overall goal was to obtain approximately three randomly selected classes within each high school to administer the YRBS and another three to administer the YHS. The number of classes per school could vary depending on expected enrollments of schools. A target was set for obtaining 3000 completed YRBS questionnaires and 3000 completed YHS questionnaires.

**B. Middle Schools**

The population under study in this survey was all sixth through eighth grade students attending a public school in Massachusetts. For middle schools, only the YHS was to be administered. Because of this, no coordination was required with the CDC and the sample design for middle schools replicated what was done in 2004, 2007, 2009, 2011, 2013, and 2015. To assist in the sample design, a file of all public schools in Massachusetts with at least one of grades six through eight was obtained from DESE. This file contained information on the reported enrollments within each grade for each school. Contact information for the principal of each school was also included.

The target number of completed questionnaires for the middle schools was also set at 3,000, as it was for the high school YHS and the high school YRBS. Information from the DESE indicated that average class size within Massachusetts middle schools was 22.7students per class. With this in mind, a sample design was proposed which would lead to a random sample of 80 middle schools participating in the study, with two classes being randomly selected from within each participating school. This would lead to approximately 45 students being selected from within each school to complete the survey for each of the 80 schools, leading to an expected total of 3,600 students. This would allow for approximately 15% of the selected students not completing the questionnaire, and still producing a yield of 3,000 completed questionnaires. The 15% reduction could come from absenteeism, parents who would not give permission for their children to complete the questionnaire, or students who refused themselves.

The DESE file indicated there were 674 schools in Massachusetts that had at least one of sixth through eighth grade. Overall, 80% of the middle school students attended a school with all of grades six through eight. In some instances, the sixth grade would be part of the elementary school or the eighth grade would be part of the high school. These elementary schools with a 6th grade and high schools with an 8th grade were eligible for selection into the sample, but only the 6th or 8th grades were eligible for selection. Using the DESE file, an initial sample of 116 middle schools was selected. These selections were made with probabilities proportionate to the reported sixth through eighth grade enrollment of the school. No stratification by region of the state was included in the design, as it was felt that the sample was large enough to support statistically valid regional estimates (i.e. county-level or combinations of counties) without stratification. (See “Part C. Additional Analytic Issues” for more information about regional estimates.) Since a school’s participation in this survey was not mandatory, it was then necessary to call principals and solicit their school’s participation. These efforts will again be discussed later.

Within the middle schools, each selected school was randomly assigned two grades, unless the school only had sixth or eighth grade. One second period class was then randomly selected from within one of the grades while one third period class was randomly selected from the other. With a targeted number of 45 students to be selected from within each of the 116 sampled middle schools, this sample design was also self weighting for middle school students. The probability of selection for a middle school student was 0.023820.

**II. Sample Recruitment and Response**

Table 1 displays the results of the school recruitment effort. It shows that 59 of the 79 high schools contacted agreed to participate in this survey. Although the state YHS/YRBS survey administration was coordinated with Boston YRBS administration, there were two schools that were part of Boston YRBS sample that only agreed to do YRBS classes but not the selected YHS classes. For the middle schools, originally 80 schools agreed to participate, but two schools reversed their decision due to changes in the school’s or town’s priorities. As a result, 78 of the 116 schools participated. The 78 schools are below the targeted number of 80 middle schools.

**Table 1: Results of Sample Selection and Recruitment of Schools**

Agreed To Refused To Total In

Participate Participate Ineligible Sample

High Schools 59\* 20\* 0 79

Middle Schools 78 38 0 116

Total Schools 137 58 0 195

\* 59 schools participated, but only 57 completed both YHS and YRBS classes.

This led to an overall cooperation rate of 74.7% of all high schools in the state sample (59/79) and 67.2% of all middle schools (78/116). The overall cooperation rate across all schools in the state sample was 70.3% (137/195). For years we have seen improvement in middle schools’ cooperation rate (54.5% in 2007, 61.6% in 2009, 69.2% in 2011, 75.0% in 2013 and 74.7% in 2015) and 67.2% represents a drop in comparison to the last two iterations. With respect to high schools, their cooperation rate was not unilinear (85.3% in 2007, 76.5% in 2009, 80.6% in 2011, 76.0% in 2013, and 74.7% in 2015).

**Table 4: Overall Data Collection Results**

Number of Number of Number of

Classes in students completed Response

Survey: in survey: enrolled: questionnaires: rate:

Middle School YHS 156 3421 3074 89.9%

High School YHS 141 2931 2524 86.1

High School YRBS 204 3977 3328 83.7

Taking the school participation rates into account, this implies an overall success rate at getting completed YHS questionnaires from high school students of 62.2% and for the YRBS students of 62.5%. The overall success rate for middle school students is 60.4%. The YHS high school rate in 2017 is slightly lower compared to 2015 (62.2% vs. 62.5), and lower than rates in 2013 (68%), 2011 (69.2%) and 2009 (66.7%). The middle school overall success rate in 2017 is much lower than the rates in 2015 and 2013 (60.4% vs. 67.1% and 68%, respectively), lower than in 2011 (62.7%), and higher than in 2009 (55.8%).

**III. Weighting and Variance Estimation Issues**

**A. Weighting**

Weights were constructed for these data using the recommended CDC protocol. Base weights reflecting the inverse of the probability of selection were initially provided by the CDC for each high school. Three adjustments to these weights were then applied. They were:

1) a school level non-response adjustment

This adjustment was computed by first dividing the sampled schools into three strata, namely small schools, medium schools and large schools. These strata were created to have approximate equal total enrollments. Therefore, more schools needed to be in the small school stratum so that the total enrollment from these small schools would be approximately equal to smaller numbers of schools from the medium and large school strata. The weight adjustment within each stratum was then simply the total enrollment from participating schools within the stratum divided by the total enrollment from all schools within the stratum regardless of whether they participated in the survey or not. The purpose of this adjustment factor was to increase the weights for students in participating schools to account for schools that did not participate. The values of these adjustment factors were 1.326 for small schools, 1.466 for medium schools and 1.225 for large schools.

2) a student non-response adjustment

This adjustment is done within class at each school. The adjustment is to increase the weights of students within a class who completed a survey to account for students within the class who did not. This adjustment was simply the sum of the weights for the students who completed the survey in the class divided by the sum of the weights for all students enrolled in the class.

3) a post stratification adjustment

This adjustment was to make the weighted sample estimates agree with known or estimated outside estimates for several demographic factors. The particular factors used in this adjustment were grade (9 through 12), gender, and race/ethnicity. For Massachusetts, the race/ethnicity weight cells were white, Hispanic, black and other.

These non-response adjusted weights are placed on the final YHS data files in a variable named PSWGT. This non-response weight can be interpreted as the number of students represented by each completed questionnaire. Weights on the file vary from a low of 51.57 to a high of 281.49. It is important to use this weight in any statistical analysis.

A second weight is also contained on the YHS data files in the variable named “SAWGT”. This weight variable simply centers the values of PSWGT so that whereas PSWGT sums across the file to the estimated total number of students in the state, SAWGT sums to the number of completed questionnaires. Estimated sample percentages will be the same if using either SAWGT or PSWGT, but PSWGT must be used for estimating population totals (e.g., the estimated number of 6th graders who smoke cigarettes). SAWGT is simply provided as many people prefer to see frequency tables based on sample sizes instead of estimated population sizes. If using the SUDAAN PROC in SAS, SPSS Complex Samples or other statistical software that can handle complex samples, PSWGT should be used.

One additional point should be made. Individual questions differ from each other in the amount of missing data. Certain questions have a higher propensity for missing data than other questions. The weights on these files cannot take into account the amount of missing data for each question. This would imply that each survey question would have to have its’ own weight. This is not practical and overly complicates analyses. If analysis runs are made to estimate population totals and PSWGT is used, these estimates may be off due to the amount of missing data for the item under consideration. A safer and more consistent way to estimate population totals is to use the best estimate of the entire population to start, estimate a population proportion, and then apply this estimated population proportion to the starting estimated total. An example can show this best. If you wanted to estimate the number of high school students who smoke cigarettes, begin with the estimated total of high school students in the state (call this N). Then use either PSWGT or SAWGT in an analysis run to estimate the percentage of high school students who smoke cigarettes (call this P). Then, the best estimate of the total number of high school students who smoke is simply (N x P). This estimate will be the same as the weighted sum, using PSWGT, of all high school students on the data file who smoke ONLY if the variable measuring smoking has NO missing data. If missing data exists, then the sum of the PSWGTs will be incorrect and smaller than (N x P). If all estimated population totals are created using the procedure recommended, then they will all compare to each other appropriately. Again, for reporting estimated population percentages, either SAWGT or PSWGT can be used to create these estimates directly. Table 6 displays estimated population totals by grade level as computed from the file provided by the DESE.

For the YHS middle school file, weight adjustments were computed in a similar fashion and the same weight variables are placed on the final YHS data files.

**Table 6: DESE Estimated Population Totals by Grade Level**

Estimated Number

of Students

Grade Level Statewide

6th grade 70,493

7th grade 71,728

8th grade 71,718

Total Middle School 213,939

9th grade 76,179

10th grade 73,209

11th grade 71,271

12th grade 69,388

Total High School 290,047

Grand Total 503,986

**B. Variance Estimation**

The sample for this study is a complex sample design due to the clustering of students within randomly selected schools. Therefore, an assumption of simple random sampling for data analysis will almost certainly lead to estimated variances and standard errors of sample statistics that are too small. This could lead to false results for any hypothesis testing performed or estimated 95% confidence intervals that are too narrow.

On the YHS data files, a variable is included named “schoolID”. This variable identifies students clustered within each school using a numeric code which cannot be used to identify any individual school. Use of this variable in statistical packages that can handle complex samples will allow for the accurate estimation of sample variances and standard errors. Statistical packages such as SUDAAN, WESTVARS, STATA, SPSS (with the Complex Samples module), and SAS (with the callable SUDAAN PROC) can all handle such designs. It is highly recommended that such packages be used for analysis purposes.

To highlight the effect that clustering can have on estimated sample standard errors, Table 7 gives estimated design effects for numerous variables. The design effects in this table are the factors which represent the degree to which clustering leads to higher estimated standard errors as compared to simple random sampling. For example, a design effect of 1.0 means the standard error is the same as would have been obtained if simple random sampling had been used. A design effect of 2.0 means estimated standard errors are twice those that would have been obtained under simple random sampling. Each variable measured in the survey can produce its’ own distinct design effect as they depend upon the amount of correlation within a school for a particular variable. One variable may be highly correlated while another may not. It should also be remembered that simple random sampling was not an option for this study as the students could only be sampled in an economical way by using the schools to get access to groups of students.

**Table 7: Estimated Design Effects**

YHS Estimated Design Effects:

Survey Variable on the 2017 YHS: High School Middle School

Q16. Number days exercised 20 minutes in last week 1.82 1.85

Q34. Ever been physically hurt by a date 1.47 1.96

Q26. Suicide attempts in past 12 months 0.92 1.38

Q41. Age at first drink of alcohol 1.78 1.52

Q42. Number of days with at least one

alcoholic drink in past 30 days 1.89 not on survey

Q47. Age at first use of marijuana 1.43 0.98

Q48. Use of marijuana in past 30 days 1.44 1.68

Q71. Number of days smoked cigarettes in

past 30 days 1.44 not on survey

Q94. Consider being over/underweight 1.02 1.27

Q96. Been examined by dentist in past year 1.51 1.89

Q32. Been cyber-bullied in the last 12 months 1.09 1.05

Q81. Used an electronic vapor product in past 1.90 1.78

30 days

As indicated in Table 7, estimated design effects for high school students vary from a low of around 0.92 for suicide attempts in past 12 months to a high of around 1.90 for using an electronic vapor product in past 30 days. Overall, average design effects for high schools seem to be around 1.48, indicating a 48% increase in estimated standard errors due to the clustering of students. For middle school students, the design effects vary from a low of 0.98 for age at first use of marijuana to a high of 1.96 for ever been physically hurt by a date. Overall, average design effects for middle schools seem to be around 1.54, indicating a 54% increase in estimated standard errors due to the clustering of students. If statistical packages that can handle complex sample designs are not available, then it is recommended that estimated standard errors be computed by first using SAWGT in SAS, SPSS, or another statistical package. The estimated standard error that results from this computation can then be multiplied by an appropriate estimated design effect from Table 7. This will inflate the value of the standard error computed in the basic computer printout to be closer to the true standard error that takes account of the complex sample design. It is important to use SAWGT for these analysis runs so as not to confuse the statistical package into believing the sample size is larger than it is since PSWGT sums to estimated population totals. For statistical packages that can handle complex sample designs, PSWGT should be used.

**C. Additional Analytic Issues**

The sample for this survey was designed to produce accurate statewide estimates by grade level. Considering the final sample sizes and the average design effects, it is recommended that subgroups of the population have a minimum of 225 sample observations before statistics should be reported for that subgroup. By grade level, each grade has far more that this. There is a variable on the file named “COUNTY” which indicates the county the school resided in. Care should be taken using this variable as some counties have small numbers of students. The number of students should be checked before reporting any statistic for county or any other population sub-group of interest.

It is also important to keep in mind that some of the newly added slang terms for drugs were confusing to students in middle school, especially those in 6th grade who raised questions regarding terms like “brown sugar”, “candy” and especially “skittles”. In addition, when asked about using inhalants, the note about what inhalants include was no longer separated as introduction; instead, it was built into the question, so some students asked whether inhalants referred to their asthma inhalers. This is important to take into consideration when analyzing the middle school students’ drug use, but also to remember when revising the future survey instruments. Future instruction about inhalants should not just say what inhalants include, but also that asthma inhalers are not included.