**YHS 2021 SURVEY METHODOLOGY**

**I. Overview**

The 2021 MYHS was conducted among 6th-12th grade students in a representative sample of public schools across Massachusetts during the Fall 2021 semester. The YHS uses two instruments, one for high school students administered primarily by web with paper and pencil (PAPI) back up, and one for middle school students administered by PAPI only. The survey took one class period to complete. Survey procedures were designed to protect student privacy and allow for anonymous participation.

The universe for the MYHS is all students attending public schools, containing any of the grades 6 through 12, in Massachusetts. Students were selected first by selecting schools, then by selecting intact classes within sampled schools. All students in selected classes were invited to participate.

The survey has historically been conducted during spring of odd-numbered years. However, due to the COVID-19 global pandemic, most schools nationwide, including those in Massachusetts, began their 2020-2021 school year with students learning remotely. In November 2020, CDC decided to delay the national YRBS until the fall of 2021 in the hopes that students nationally would return to in-person learning at the start of the 2021-2022 school year. Following suit, most states, including Massachusetts, also decided to defer their state YRBS administration until the fall of 2021. Since the MYHS and YRBS are co-implemented, the YHS timeline was altered to reflect the changes In the YRBS timelines.

District and school recruitment for a fall administration began in May 2021. The fielding window for survey administration was October 4, 2021 to December 22, 2021.

**II. Sample Design**

**A. Middle School Sample**

Middle schools were selected from the sampling frame (provided by DESE) with probability proportional to the number of students in grades 6-8 attending the school. Based on an expected middle school response rate of 68%, 117 schools were selected with a target of 79 responding schools. The school selection probability for school i was calculated as P\_i=N\_i/N n\_i , where Ni is the number of enrolled students in the eligible grades in school i, N is the total number of students enrolled in eligible schools in grades 6-8 in the state, and ni is the number of schools to be selected.

During recruitment, responding schools were asked to provide a complete list of classes covering all students in eligible grade levels in a required class, (e.g., English) or a fixed period of the day (e.g., 2nd period) for class selection. ICF reviewed the list to remove any classes that did not meet eligibility criteria (e.g., classes meeting off campus, intact classes of low-literacy readers or physically/cognitively impaired students unable to complete the survey independently). Approximately two classes in each selected middle school were sampled. Classes were selected by calculating a sampling interval equal to the total number of classes meeting the selection criteria (a required course or period of the day) expected to be offered based on the school’s enrollment on the frame, divided by the desired number of classes to be selected (two). Classes on the list provided by responding schools were assigned an ID number sequentially. A random starting point was assigned to each school, and the classes with ID numbers equal to a multiple of the sampling interval plus the starting point were selected. The selection probability for class c within school i was calculated as P(c\_i )=2/C\_i where Ci is the expected number of eligible classes in school i.

**B. High School Sample**

By design, the MYHS high school sample had 100% overlap with schools selected for the YRBS by CDC's TA Contractor. MYHS class interval numbers were drawn by CDC’s TA Contractor at the time that YRBS class intervals were drawn. Based on an expected high school response rate of 76%, 79 schools were selected with a target of 60 responding schools. High school class selection followed the same procedures described above for middle schools, however classes for both the YRBS and MYHS were selected concurrently from the same class list using interval numbers unique to each study. That is, a high school class could be selected for the MYHS or the YRBS, but not both. For the MYHS administration in high schools, approximately three classes consisting of an average of 58 total students were selected per participating school.

The final MYHS sample included 139 school districts, 116 eligible middle schools, and 79 high schools. At the middle school level, the two-stage sampling strategy had a design effect (DEFF) of 1.98, resulting in a margin of error of +/- 2.5%. At the high school level, the DEFF was 2.14 with a margin of error of +/-2.6%.

**C. Sample Validation**

Prior to the initiation of study invitation, the MYHS sample was validated by ICF staff to ensure school eligibility and confirm district and school leadership. When recruitment and fielding occur within the same school year, there are typically few changes in district and school leadership encountered. However, the delay in fielding the 2021 MYHS from spring to fall shifted recruitment activities to span two consecutive school years. Thus, it was necessary to have an additional inquiry regarding anticipated changes in district or school leadership and school grade composition for the following year. In addition, districts and schools were asked to verify their physical address and schools were asked to confirm school enrollment. During validation, one school in the middle school sample was found to be ineligible as it served grades one through four, making the final middle school sample 116 schools. All 79 schools in the high school sample met eligibility requirements.

**III. Sample Recruitment and Response**

At the school level, 196 schools were selected across Massachusetts to participate in the MYHS, 116 middle schools and 79 high schools. In total, 68.10% (n= 79) of selected middle schools and 75.95% (n=60) of selected high schools participated in the study. The remaining 24.05% (n=19) of selected schools were considered refusals. Of those refusals, 47.37% (n=9) of them were due to district-level refusals to allow contact with schools to discuss participation and 52.63% (n=10) were school level refusals.

Of the 60 high schools that participated, 16.67% (n=7) used the paper-based survey and the remaining 88.33% (n=53) opted for students to participate using the web-based survey. As noted above, the middle school survey was administered via PAPI only.

**Table 1: School Participation Rates, by School Level and Mode**

|  |  |  |  |
| --- | --- | --- | --- |
| School Level | Sampled Schools | Participating Schools | Mode |
| **Web** | **Paper** |
|  | **#** | **#** | **%** | **#** | **%** | **#** | **%** |
| Middle Schools | 116 | 79 | 68.10% | 0 | 0% | 79 | 100% |
| High Schools | 79 | 60 | 75.95% | 53 | 88.33% | 7 | 11.67% |

The student participation rate for the 2021 MYHS was 81.38% for high school students and 86.00% for middle school students. Overall, 3,260 eligible students from the 60 participating high schools and 3,344 eligible students from the 79 participating middle schools were invited to participate in the survey, and 2,645 high school students and 2,876 middle school students did so.

Of the 2,645 high school students that participated, 8.77% (n=232) of them completed the paper-based survey. The remaining 91.23% (n=2,413) were captured by the web-based survey.

##### **Table 2: Student Participation Rates, by School Level and Mode**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **# Eligible** | **# Web Completes** | **# PAPI Completes** | **Total Completes** | **Participation %** |
| **Middle School students** | 3,344 | 0 | 2,876 | 2,876 | 81.38% |
| **High School Students** | 3,260 | 2,413 | 232 | 2,645 | 86.00% |

Taking the school and student participation rates into account, this implies an overall response rate for high school students of 65.32%. The overall response rate for middle school students is 55.42%.

**IV. Weighting and Variance Estimation Issues**

**A. Weighting**

The MYHS sample design was a probability sample, which allowed for the calculation of sampling weights. The multistage sample was designed so that each student at a given level had the same overall probability of selection. The design yielded unbiased population estimates because the probability of selection was known for every student.

Similar to CDC’s weighting plan for the YRBS, the weighting plan for the MYHS followed the sample design with adjustments for nonresponse at each stage of sampling and a final adjustment to known middle school and high school population totals by grade, race/ethnicity, and sex provided by DESE. We adjusted for nonresponse at each stage of weighting to minimize nonresponse bias. Nonresponse bias occurred with the presence of both of the following conditions:

1. Nonrespondents differed from respondents along key survey characteristics and

2. Nonrespondents accounted for a large enough proportion of the sample (i.e., substantial nonresponse rates).

While nonresponse adjustments reduce bias, they usually introduce added variation to the weights. A balance between bias reduction and the increase in variance were considered when implementing a nonresponse adjustment. Nonresponse adjustments use the information available for the sampled cases. In general, the adjustment distributes the base weights of the nonrespondents to the responding sampled cases so that the sum of the adjusted weights equals the sum of the base weights.

The weighting plan for both the middle school and high school surveys replicated the methodology used by CDC on the national YRBS. The overall weights were decomposed as the products of successive stage weights and weight adjustments (the latter adjustments are denoted by a hat symbol). The middle school and high school MYHS data were weighted separately but followed the same weighting process. The middle school weights were computed using the sampling probabilities developed by ICF. The high school weights used the school and class probabilities provided by CDC from the coordinated sample draw with the YRBS.

**School Sampling Weight (W1)**

The weighting process began with W1 computed as the inverse of the school probability of selection ($p\_{i}$).

$$W1= \frac{1}{p\_{i}}$$

**School Nonresponse-Adjusted Weight (W2)**

We then adjusted for school nonresponse within weight adjustment cells defined by school level and school size. Schools were categorized according to enrollment as small, medium, or large with an approximately equal number of schools per category.

The nonresponse adjustment factor ($\hat{W}1$) was calculated as as the ratio of two weight sums: the sum of W1 for all sampled schools (A) divided by the sum of W1 for participating schools (P). The formula for the school nonresponse adjustment factor $\hat{W}1$ is:

$$\hat{W}1= \frac{\sum\_{A}^{}W1}{\sum\_{P}^{}W1}$$

W2 was computed as the product of the school weight and the school nonresponse adjustment factor:

$$W2= W1∙\hat{W1}$$

**Class Weight (W3)**

The class weight was calculated as the product of two factors: (1) the nonresponse-adjusted school weight and (2) the class sampling weight, computed as the inverse of the probability of selecting a class J within a responding school. Specifically,

$$W3= W2\frac{1}{p\_{J}}$$

**Class Nonresponse-Adjusted Weight (W4)**

Within a school, a class nonresponse adjustment factor ($\hat{W}3$) was calculated as the number of selected classes (n) over the number of participating classes in the survey (k).

$$\hat{W}3= \frac{n}{k}$$

We expressed W4 as follows:

$$W4= W3∙\hat{W3}$$

Student Weight (W5)

Within each classroom, W5 approximately accounted for student nonresponse within a class. We calculated a student factor by dividing the class enrollment (a), obtained during the data collection process, by the number of responding students ($b$).

$$\hat{W}4= \frac{a}{b}$$

We expressed W5 as follows:

$$W5= W4∙\hat{W4}$$

**Final State Sample Student Weight (W6)**

The student weights were post-stratified to population totals on the sampling frame by grade (x), gender (y), and race/ethnicity (z) to ensure the weighted population estimates match known state population totals. Within each grade by gender and by race combination, the post-stratification adjustment factor was defined as follows:

$$\hat{W}5= \frac{N\_{xyz}}{\sum\_{xyz}^{}W5}$$

The post-stratum cell totals are the population totals $N\_{xyz}$ known for each cell defined by grade (x), gender (y), and race/ethnicity (z) categories.

W6 was computed as the product of the student weight and the post-stratification adjustment:

$$W6= W5∙\hat{W5}$$

**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.