Sample Description and Methodology

Sample of Schools

School Type	Middle School	High School
Regular	Included	Included
Charter	Included	Included
Private	Excluded	Excluded
Alternative	Excluded	Excluded
Vocational	Included	Included
Special Education	Excluded	Excluded

Methodology

For a complete copy of the CDC YRBS Methodology, follow the link below:
http://www.cdc.gov/mmwr/pdf/rr/rr6201.pdf

Sampling, Weighting, and Response Rates

A two-stage, cluster sample design is used to produce a representative sample of students in grades 6-8 and 9–12 in Maryland, stratified for Maryland's 23 counties and Baltimore City (to ensure adequate sample sizes[1] for local estimates). This design produces approximately 50,000 useable answer sheets for high school statewide and 35,000 for middle school.

In the first sampling stage, schools are selected with probability proportional to school enrollment size (enrollment in eligible grades). In Maryland, for some local school systems, these procedures are modified to meet the individual needs of the sites. For example, in a given site, all schools, rather than a sample of schools, might be selected to participate.

In the second sampling stage, intact classes of a required subject or intact classes during a required period (e.g., second period) are selected randomly.

All students in sampled classes are eligible to participate (cluster).

Those surveys that have a sample selected according to the protocol described above, appropriate documentation of school and classroom selection, and an overall response

[1] Maryland instructs the CDC to draw samples at a 95% confidence level that will produce local school system estimates of the prevalence of risk behaviors at ±3% for grades 9-12 (high school), and ±5% for grades 6-8 (middle school).

rate of ≥60% are weighted. These three criteria are used to ensure that the data are representative of students in grades 6-8 and 9–12 in that jurisdiction.

The overall response rate is calculated by multiplying the school response rate by the student response rate.

A weight is applied to each student record to adjust for student nonresponse and the distribution of students by grade, sex, and race/ethnicity in each jurisdiction. Therefore, weighted estimates are representative of all students in grades 6-8 and 9–12 in each jurisdiction.

Surveys that do not have an overall response rate of ≥60% and appropriate documentation are not weighted. Unweighted data represent only the students participating in the survey. Among the state surveys, school response rates ranged from 73% to 100%, student response rates ranged from 60% to 88%, and overall response rates ranged from 60% to 84%.

Data-Collection Protocols

Trained data collectors employed by Maryland's survey contractor travel to each selected school to administer the questionnaires to students. These data collectors read a standardized script to participating students. The script includes an introduction to the survey. Data collectors also record information about schools and classrooms (e.g., grade level of classes sampled and number of students enrolled in a sampled class). This information is used later in the survey process to verify sample selection and to weight data.

Procedures for survey are designed to protect student privacy by allowing for anonymous and voluntary participation. Students complete the self-administered questionnaire during one class period and record their responses directly in a computer-scannable answer sheet. To the extent possible, students' desks are spread throughout the classroom to minimize the chance that students can see each other's responses, or if taking the survey in an environment where several classes are present, procedures utilized by the school for administering standardized tests or assessments in such a setting are utilized.

Upon completion of the survey, answer sheets are gathered and placed into manila envelopes which are then sealed.

Students who are absent on the day of data collection still can complete questionnaires if their privacy can be maintained if the school and classroom teachers permit. These make-up data-collection efforts are performed by school personnel. Students making up the survey are further supplied with an envelope in which they seal their answer

sheet prior to turning it in to school personnel. School personnel are required to forward the sealed envelope(s) to the DHMH survey contractor. Allowing students who were absent on the day of data collection to take the survey at a later date increases student response rates. In addition, because absent students, especially those who are absent without parental permission, are more likely to engage in health-risk behaviors than students who are not absent, make-up data collection procedures help provide data representative of all high school students.

Data-Processing Procedures

Data processing is a collaborative effort between CDC and its technical assistance contractor (Westat) that provides a system of checks and balances. Maryland's survey contractor sends completed answer sheets to the CDC contractor, which scans them and constructs a raw electronic dataset. The CDC contractor sends all raw datasets to CDC, which edits them to identify out-of-range responses, logical inconsistencies, and missing data. The data cleaning and editing process is performed by the Survey Data Management System (SDMS), which CDC developed in 1999 and updated in 2008 to a web -based system in 2008 and performs its functions using Visual Basic, SAS, and SUDAAN programs. The processing system accommodates questionnaires in which questions have been deleted or added by the sites by first screening them to note differences from the standard questionnaire and then accounting for those differences during processing.

Approximately 180 logical edits are performed on each standard questionnaire. Responses that conflict in logical terms are both set to missing, and data are not imputed. For example, if a student responds to one question that he or she has never smoked but then responds to a subsequent question that he or she has smoked two cigarettes during the previous 30 days, the processing system sets both responses to missing. Neither response is assumed to be the correct response.

Questionnaires with <20 valid responses remaining after editing are deleted from the dataset. In 2011, the median number of completed questionnaires in the state surveys that failed quality-control checks and were entirely excluded from analysis was 13 (range: 0–351).

Additional data edits are applied to the height, weight, and BMI variables to ensure that the results are biologically plausible. These three variables are set to missing when an observation lies outside logical limits developed by CDC's Division of Nutrition, Physical Activity, and Obesity. In 2011, the median number of completed questionnaires in the state surveys that had their height and weight data set to missing because either these values or their resulting BMIs were considered implausible for the student's sex and age was 32 (range: 7–567).

Edited data are sent to the CDC contractor for weighting. If response rates are sufficient, documentation is complete, and the site followed sampling protocols correctly, the contractor weights the data according to the procedures previously described in this report and sends the weights to CDC, which merges the weights onto the edited data file. CDC and Westat use file transfer and tracking functions built into the Survey Technical Assistance Website to ensure that all transfers are logged and reported.

Reports and Publications

CDC generates a report for each participating site. Before 2013, each report contained approximately 500 pages in a three-ring binder divided into two sections: survey results and survey documentation.

The survey results section included a sample description; bar charts and pictographs summarizing key results; tables and graphs that provided prevalence estimates for each question, including site-added questions; and a report that provided the results of trend analyses using logistic regression to test whether results have changed over time. The tables provided estimates overall and by sex, race/ethnicity, grade, and age and included 95% confidence intervals for all sites with weighted data. To help ensure the reliability of the estimates and protect the anonymity of respondents, subgroup results were not reported if any subgroup contained <100 students. The graphs provided estimates overall and by sex, grade, and race/ethnicity, and were provided as PowerPoint files to facilitate presentation of results. CDC used SDMS to generate these reports with customized Visual Basic programs that ran SAS, SUDAAN, and Crystal Reports to generate multiple output files. The survey documentation section of the report included a copy of the site's questionnaire, a data user's guide describing how data were edited and how each variable was calculated, a codebook for the electronic data set, information on sampling and weighting, a Sample Statistics Report including the standard error and design effect for each variable, and additional resource documents such as "Understanding, Analyzing, and Presenting Your YRBS Data," "How to Interpret YRBS Trend Data," and "Software for Analyzing YRBS Data." In addition, each report contained a CD-ROM with an electronic copy of the site's data in multiple data file formats (e.g., SAS, SPSS [*41*], and ASCII) to permit subsequent analyses and an electronic copy of all the material described above.

Beginning with the 2013 cycle, CDC uses the SAS Output Delivery System instead of Crystal Reports as part of SDMS. **To reduce environmental impact and cost, all report materials described above will be provided in electronic format only on a CD-ROM**. Sites have been using electronic files to share their data and results with others and to post results on their websites, and they will continue to be able to do so.

To ensure the timeliness of the data, CDC typically completes site reports within 12 weeks of receipt of completed questionnaires or answer sheets. Because surveys generally are completed in the spring, most sites receive their reports during the summer, so they can use their survey results to help plan for the coming school year.

Data Quality

From the inception of YRBSS, CDC has been committed to ensuring that the YRBS data are of the highest quality. Obtaining high-quality data begins with high-quality questions.

The original questionnaire was subjected to laboratory and field testing, and CDC conducted reliability testing of the 1991 and 1999 versions of the questionnaire. In addition, two studies have been conducted to assess the effect of implementing changes to the questions that assess race and ethnicity. CDC made these changes to comply with new standards established by the Office of Management and Budget in 1997. The first study tested the effect of changing the question from one in which students were required to select a single response to one that allowed students to select one or more responses. The second study tested the effect of changing from a single-question format that asked about race and ethnicity together to a two-question format that asked separate questions about race and ethnicity. Both studies indicated that the changes to the questions had only a minimal effect on reported race/ethnicity and that trend analyses that included white, black, and Hispanic subgroups were not affected.

Another aspect of data quality is the level of nonresponse to questions. For the 2011 national YRBS, nonresponse attributed to blank responses, invalid responses, out-of-range responses, and responses that did not meet edit criteria ranged from 0.5% for the question that assesses the sex of the respondent to 14% for the question that assesses the race of the respondent. For 91% of all questions, the nonresponse rate was <5%, and for 16% of all questions, the nonresponse rate was <1%.

To further ensure data quality, survey administrators use standardized procedures. To determine how using different procedures can affect survey results, CDC has conducted a series of methods studies. In the first study, conducted in 2002, CDC examined how prevalence estimates were affected by varying honesty appeals, the wording of questions, and data-editing protocols, while holding population, setting, questionnaire context, and mode of administration constant. The study indicated that different honesty appeals and data-editing protocols did not have a statistically significant effect on prevalence estimates. In addition, the study indicated that, although differences in the wording of questions can create statistically significant differences in certain prevalence estimates, no particular type of wording consistently produced higher or lower estimates.

In 2004, CDC conducted a study to determine how varying the mode and setting of survey administration might affect prevalence estimates. While previous research had examined the effects of varying setting (school versus home) and mode (paper-and-pencil instrument [PAPI] versus computer-assisted self-interview [CASI]), this study was the first to assign school classes randomly to one of four conditions in which mode and setting were varied systematically: school-based administration using PAPI, school-based administration using CASI, home-based PAPI administration, and home-based CASI administration. Results revealed that students completing questionnaires at school were more likely to report health-risk behaviors than students completing questionnaires at home, but that mode effects were weaker: prevalence estimates for health-risk behaviors generally did not differ for CASI and PAPI administrations, and these effects were independent of setting. On the basis of these results, CDC decided to continue with PAPI administration for the YRBSS. Because the use of CASI did not increase the reporting of risk behaviors, its increased cost and complicated logistics did not appear to be justified.

In 2008, CDC conducted two additional studies to determine the feasibility and effect of conducting YRBS as a web-based survey. In the first study, classes in grades 9 and 10 were assigned randomly to complete the YRBS in three conditions: 1) using PAPI in the classroom, 2) using web-based CASI administration in school computer labs or in classrooms with sufficient numbers of computers, or 3) using web-based CASI that students could complete on their own (i.e., at any time at any computer with Internet access). Results indicated that risk behavior prevalence estimates generated from PAPI and CASI administered in a classroom setting in schools generally were equivalent. However, web-based CASI administration yielded more missing data than PAPI administration, and web-based "on their own" administration yielded an unacceptably low response rate. In addition, perceived and actual privacy and perceived anonymity were compromised when administering in-class web-based questionnaires.

In the second study, paper-and-pencil questionnaires were mailed to a nationally representative sample of public and private high school principals to assess computer availability in U.S. schools and to assess principals' perceptions of web-based student surveys. Although 64% of principals preferred web-based student surveys to those conducted via PAPI, only 30% said they would be more likely to agree to participate in such a survey if it were conducted online. Further, this study revealed that many schools do not have sufficient computer capacity to participate in an in-class web-based survey. As a result, web-based student surveys could create a significant burden on schools and lead to unacceptably low school participation rates. Taken together, the results of the 2008 studies informed CDC's decision not to convert the YRBSS from PAPI to web-based administration so the quality of the system could be maintained.

Limitations

The YRBS is subject to at least five limitations. First, all YRBS data are self-reported, and the extent of under-reporting or over-reporting of behaviors cannot be determined, although studies described in this report demonstrate that the data are of acceptable quality. Second, the school-based data apply only to youths who attend school and therefore are not representative of all persons in this age group. Nationwide, in 2009, approximately 4% of persons aged 16–17 years were not enrolled in a high school program and had not completed high school. The NHIS and Youth Risk Behavior Supplement conducted in 1992 demonstrated that out-of-school youths are more likely than youths attending school to engage in the majority of health-risk behaviors. Third, local parental permission procedures are not consistent across school-based survey sites. Fourth, state-level data are not available for all 50 states. Three states (Minnesota, Oregon, and Washington) do not participate, and in 2011, four states (California, Missouri, Nevada, and Pennsylvania) did not obtain weighted data. Finally, YRBSS addresses only those behaviors that contribute to the leading causes of morbidity and mortality among youths and adults. However, school and community interventions should focus not only on behaviors but also on the determinants of those behaviors.