# YRBSS Technical Assistance Memo

## Overview

The Youth Risk Behavior Surveillance System (YRBSS) monitors the engagement of youth in risky activities as…

* “Behaviors that contribute to unintentional injuries and violence”
* “Sexual behaviors related to unintended pregnancy and sexually transmitted diseases, including HIV infection”
* “Alcohol and other drug use”
* “Tobacco use”
* “Unhealthy dietary behaviors”
* “Inadequate physical activity”

“YRBSS also measures the prevalence of obesity and asthma and other priority health-related behaviors plus sexual identity and sex of sexual contacts.” [[1]](#footnote-1)

Within this system, the primary source of publicly-available data for consistent trend analysis is from the National Youth Behavior Survey (YRBS) conducted by ICF Macro, Inc. under contract to the U.S. Department of Health and Human Services (HHS), Centers for Disease Control and Prevention (CDC). This is a biennial survey of high school students. Additionally, CDC funds state- and local-level surveys (using a similar instrument), for which data, if publically available, must be directly acquired from the sponsoring agencies (i.e., of state and local governments). There have also been specialized surveys conducted on a one time basis. However, within this memo, we limit our discussion to the methods and use of data from this series (i.e., National YRBS).

## Sample Design

The Youth Risk Behavior Survey has been conducted on a biennial basis since 1994. Over the course of its fielding, sample designs have changed to some degree. However, it has always been organized as a three-stage cluster sample of high school students:[[2]](#footnote-2)

**First Stage:** the Primary Sampling Units (PSUs) are large counties or clusters of smaller counties. These clusters are organized and sampled by strata. Stratification is according to urban/rural status and the percentage of black and Hispanic students. Since the 1999 sample, PSUs large enough to be selected with certainty are divided into sub-PSU units. Schools then are sorted by size and assigned in rotation to the newly created sub-PSU units.

**Second Stage:** the sampling units are schools, within each sampled PSU, schools are selected systematically on a probability proportionate to size (PPS) basis. The size measure is essentially the number of enrolled students, but black and Hispanic students are counted with higher weight in order to allow for more precise estimates for those populations. Among schools sampled, some choose not to participate in the survey.

**Third Stage:** From each participating, sampled school, students are organized into distinct groups based on their assigned class for a specific subject such as English or social studies. For each sampled school, one or two classes are selected for surveying. Within these classes, all students are provided the opportunity to complete the survey questionnaire, if they choose.

In Table 1 below (see Page 5), we provide survey fielding statistics by year.

## Oversampling

To allow the survey to make estimates specific for blacks and Hispanics at a desired rate of precision, several oversampling strategies are used:

1. For Stage 1: higher sampling rates in strata with higher percentages of blacks and Hispanics.
2. For Stage 2: modified size measure in PPS computation. Essentially, when computing the size measure applied in sampling, black and Hispanic students count for more than other students.
3. For Stage 3: schools with high minority enrollment have two classes sampled, otherwise only one.

The result of oversampling is that members of the targeting groups have a higher probability of selection into the sample than they would otherwise. Since the survey weight attached to each respondent record is intended to reflect the number of students whom it represents, the weights for members of targeting groups are lower than non-targeted groups directly in proportion to the inverse of the sampling rate. As a result, estimates made using these weights will be unbiased. Also, by using survey estimation software, the estimation of variance will appropriately account for the oversampling in the context of the complex survey design. Nevertheless, users should be cognizant that where oversampling has been performed, estimates can be made more precisely (i.e., with lower standard errors) than would otherwise be possible based on the size of their population.

## Estimation

Because of the sample design is complex (i.e., it uses multi-stage clustering and non-equal probability sampling), estimation of point estimates and their error requires the use of appropriate survey estimation software. Unlike, the National Survey of Drug Use and Health (NSDUH) there is no overlap in cluster selection from year to year. Thus, estimates from each years’ survey can be computed independently and then compared. For yearly comparisons estimates, statistical testing requires the estimation of the standard errors of the year-to-year differences. Generally, because of their statistical independence (because they derive from independently drawn random samples), these can be estimated as

$$SE\_{Diff}=\sqrt{SE\_{1}^{2}+ SE\_{2}^{2}}$$

…where

$SE\_{Diff}$ - is the estimated standard error of the difference between the two estimates

$SE\_{1}$ - is the estimated standard error for estimate 1

$SE\_{2}$ - is the estimated standard error for estimate 2

Alternatively, if the user wishes to have these comparisons computed directly from available survey estimation procedures, they can do this by appending the yearly files together prior to analyzing them together. There is existing Centers for Disease Control and Prevention (CDC) documentation that show how to properly specify the sample design and then make estimates in commonly used statistical packages:

<https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2015/2015_yrbs_analysis_software.pdf>.

At a minimum, software packages will require description of the survey design with the following pieces of information:

* Stratification: specified by the variable <STRATUM>
* Clustering: specified by the variable <PSU>
* Weighting: specified by the variable <Weight>

Additionally, users should be careful when making estimates for sub-populations (also called domains). For these, it is necessary that all responses are processed within the software and the use of domains be specified explicitly in the coding for the procedure run. For example, that if a user is making estimates for black females, they should **not** remove records for other respondents (i.e., those not black and female) prior to having survey data processed for estimation by the procedure. In this case, a dummy variable can be created that flags responses of interest (e.g., **Flag\_Analysis\_Group** = **1** if black and female, **Flag\_Analysis\_Group** = **0**, otherwise) and then use this dummy variable (**Flag\_Analysis\_Group)** to distinguish those responses in the domain of interest from the others.

Code provided for analysis of NSDUH could then be used with slight modifications. Because YRBS uses different variable names than NSDUH for its survey design variables, the following variable name transformations should be made:

* STRATUM 🡪 vestr
* PSU 🡪 verep
* Weight 🡪 analwt\_c

Additionally, analysis and level variables should be converted into numeric format, and the survey year should be specified numerically in the variable Year, as shown in the example SAS code segment below. This code generates an analysis file usable within the NSDUH estimation macros:

**data** Analysis; \*\* 🡨Output: File used for analysis;

set YRBS2015; \*\* 🡨 Input: YRBS Public Use File;

Year = **2015**; \*\* Set year value;

RaceEth\_Num = input(RaceEth, **2.**); \*\* Convert into numeric variable;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

\*\* Rename survey design variables:

rename STRATUM = VESTR

 PSU = VEREP

 WEIGHT = ANALWT\_C;

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*;

**run**;

Generally there are several types of estimation methods, such as Taylor series linearization (TSL), Balanced Repeated Replication or Fay’s Method, and Jackknife. These can be specified when applying each software procedure. In most cases, the methods “give comparable variance estimates and neither is clearly preferred.”[[3]](#footnote-3) For most purposes, the use of Taylor series is most straightforward and is the default method for software.[[4]](#footnote-4) The exceptions to this are when domains are small (i.e., when trying to make estimates for a small group, such as Native Americans, or estimating medians and especially more extreme order statistics, such as 95 percentile. Here TSL can be expected to underestimate variance and jackknife can produce more accurate (albeit more computationally intensive) estimates.[[5]](#footnote-5),[[6]](#footnote-6)

More detail about the use of these in various software systems is provided in this document: <https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2015/2015_yrbs_analysis_software.pdf>

## Response Variables

As is typical with surveys administered repeatedly over a number of years, questionnaires change over time and for any given response variable (assuming it is collected with more-or-less the same question), its values may not be consistently coded from year-to-year. For this reason it is incumbent on the data user to review available documentation: <https://www.cdc.gov/healthyyouth/data/yrbs/data.htm> (Under National YRBS Datasets and Documentation, Data User’s Guide) specific for each year for each response variable to be analyzed. In cases where the mapping of responses from year to year is straightforward, the data user can apply this mapping to the files to establish consistency prior to making pooled estimates or year-to-year comparisons, or conducting trend analysis.

|  |
| --- |
| **Table 1. Summary of National Youth Behavior Risk Survey Sampling** |
|   | **Survey Year** |
| **Statistics by Sampling Stage** | **2015** | **2013** | **2011** | **2009** | **2007** | **2005** | **2003** | **2001** | **1999** | **1997** | **1995** | **1993** |
| **Stage 1** |   |   |   |   |   |   |   |   |   |   |   |   |
| **A.** Number of PSUs in Frame | N/A | N/A | N/A | N/A | N/A | N/A | 1,262 | 1,256 | 1,270 | 1,719 | 1,955 | 1,928 |
| **B.** Number of PSUs Selected | N/A | N/A | N/A | N/A | N/A | N/A | 57 | 57 | 52 | 54 | 52 | 50 |
|  |   |  |  |  |  |  |  |  |  |  |  |  |
| **Stage 2** |   |   |   |   |   |   |   |   |   |   |   |   |
| **C.** Number of Schools Selected in sample | 180 | 193 | 194 | 196 | 195 | 203 | 195 | 199 | 187 | 191 | 157 | 199 |
| **D.** Number of Schools Participating | 125 | 148 | 158 | 158 | 157 | 159 | 158 | 150 | 144 | 151 | 110 | 155 |
| **E.** School Response Rate | 69% | 77% | 81% | 81% | 81% | 78% | 81% | 75% | 77% | 79% | 70% | 78% |
|  |   |  |  |  |  |  |  |  |  |  |  |  |
| **Stage 3** |   |   |   |   |   |   |   |   |   |   |   |   |
| **F.** Student Response Rate | 86% | 88% | 87% | 88% | 84% | 86% | 83% | 83% | 86% | 87% | 86% | 90% |
| **G.** Number of Students in Sample**\*** | 18,165  | 15,480  | 17,672  | 18,573  | 16,662  | 16,262  | ***18,330***  | ***19,757***  | ***17,848***  | ***18,691***  | ***12,679***  | ***18,107***  |
| **H.** Number of Questionnaires Submitted by Students (before data editing) | 15,713  | 13,633  | 15,503  | 16,460  | 14,103  | 13,953  | N/A | N/A | N/A | N/A | N/A | N/A |
| **I.** Number of Usable Questionnaires Submitted by Students (after data editing) | 15,624  | 13,583  | 15,425  | 16,410  | 14,041  | 13,917  | 15,214  | 16,398  | 15,349  | 16,262  | 10,904  | 16,296  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |
| \*The ***italicized*** values in Row **G** (for **1993** to **2003**) are NORC-estimates computed by dividing the value in Row **I** (Usable Questionnaires) by Row **F** (Student Response Rate), which are reported in the Data User’s Guides for each survey year. |

## Missing Values

In the process of developing translation tables for YRBSS, it was discovered that for some questions, in certain years there were very high levels of responses being missing (See Appendix A, National YRBSS Missing Data for a fuller discussion of the presence of variables with high levels of missingness). This is related to the conduct of YRBSS sampling. YRBSS is sampled at the national level, but is also resampled selectively at the state or local levels. Because the questionnaires associated with these samples are not completely identical, when the same school has been included in two samples, they may return the non-national level questionnaire (which exclude certain questions) for the national sample. This can lead to estimation issues for variables having substantial levels of missingness.

When making survey estimates, statistical software generally excludes records with relevant missing values from the computation, whether the missing value is in the response, class-level, or weight variables.[[7]](#footnote-7) This implicitly assumes that the occurrence of missing values among the records occurs independently of the actual values they represent. This situation is commonly called, completely missing at random. On the other hand, it is generally the case that missingness is patterned. For example, if a question is asking about sexual activity, it may be the case that females are less willing to answer than males, and therefore their responses have more missing values. This type of missingness (i.e., being missing not completely at random) has implications for both point estimates and variance estimates.

For the point estimates, if for any group, missingness is correlated with the measure being summarized, the resulting estimate will be biased: the expected value of that estimate will not be equal to the true value. For example, if respondents who have used a certain type of drug are more likely not to answer questions about having used it, then the overall estimates of drug use would be biased low: the estimate will likely fall under the true value of mean usage. As for variance estimates, generally if we were to assume that missing values occur completely at random this would imply a lower variance than if this were not the case. Thus, if we estimate variance assuming missing completely at random this will tend to underestimate true variance if this were not the case. To some degree, statistical software can be adjusted so that missing completely at random is not assumed. In SAS this is done by setting the <NOMCAR> option in <PROC SURVEYFREQ>, <PROC SURVEYMEANS>, or <PROC SURVEYREG>. It should be noted that invoking this option will not affect the point estimate, which may be of more concern to the analyst. Here the proper adjustment requires the proper application of an imputation technique.

More to the point, for response variables in YRBSS having a high level of missingness the accuracy of derived estimates should be carefully scrutinized. For example, for 2001, the variable <ECSTASYFRQEV> or <Q90> on the public use file was only collected on the national YRBSS and so it would be expected to have a high level of missingness. In fact, the response is missing 2,303 times among the 13,601 responses included on the public use file. Because the missingness will be geographically clustered in areas which were sampled at the state and local level, this will bias the resulting estimates. We can see these rates of missing values by region (see Table 2 below).

|  |
| --- |
| **Table 2: Reporting of Ecstasy Use on YRBSS 2001** |
| **Region** | **Q90 (Ecstasy one or more time)****Percent of Responses with Missing Values** | **Percent Reporting Never Used Ecstasy** |
| Northeast | 19.8% | 89.5 |
| Midwest | 12.8% | 90.5 |
| South | 12.1% | 89.9 |
| West | 25.0% | 87.3 |

From this we see that in the West, 25% of responses for this question were missing, which is substantially higher than in the other regions, but at the same time the percentage reported to have never used ecstasy was lower. If the missing respondent in the West had the same level of “Never Used Ecstasy” as completed respondents in the West, then the overall rate of “Never Used Ecstasy” will be over-estimated by survey. The data user should at least be cognizant of the possible biases introduced by high levels of missingness and they should consider whether they want develop an imputation model to address this issue.

## Appendix A: National YRBSS Missing Data

### Introduction

The Youth Risk Behavior Surveillance System (YRBSS) survey has two components: a national school-based survey conducted by the Centers for Disease Control and Prevention (CDC); and state, territorial, tribal, and local versions of the survey conducted by state departments of health and/or education. The National YRBSS results are NOT the aggregation of results from smaller geographic areas; the national survey data are a separately drawn sample. Both components of the survey follow the same methodology and use the same core questionnaire. State/local/tribal versions of the YRBSS may include context-specific questions and the national YRBSS typically includes several questions that are not included in the core questionnaire common to both components. The IPUMS team only harmonized the national high school public use data, as stipulated in the request for task order proposal and corresponding task order documentation.

### Missing responses in the national survey data

During the harmonization process, the IPUMS team discovered high rates of missing responses for variables that are only available on the national version of the YRBSS survey. It appears that some students selected as part of the national YRBSS sample instead complete state/local/tribal versions of the questionnaire that do not include the additional questions from the national survey. The CDC methodology reports and national user guides do not specify how or why students receive state/local/tribal versions of the questionnaire when selected into the national survey, but do provide two pieces of information that corroborate that high rates of missing are the result of students selected into the national sample receiving a state/local/tribal questionnaire.

First, the 2005 national user guide explicitly notes that 2,212 students completed a state or local YRBS questionnaire instead of the National YRBS questionnaire. In 2005, the national YRBSS modified the structure of the race question to ask about race and Hispanic ethnicity separately, while the state/local/tribal versions of the questionnaire retained the single-race question from the 2003 survey. As noted in the YRBSS translation tables deliverable complex coversheet, there are an elevated number of missing responses to the variables RACE, and HISPETH, for respondents who have valid data in the variables RACEETHORIG05 and RACEETH. Respondents with missing data in the national version of the race and ethnicity variables also have missing data for other questions asked only on the national survey in 2005.

Secondly, YRBSS methodology reports contain information explaining how students might receive a state/local/tribal version of the questionnaire even if they were sampled to be included in the national dataset. The 2004 and 2013 YRBSS Methodology reports note that schools may be sampled in both the national sample and a state/local/tribal sample (or even as part of both a state sample and a local sample). When this happens, students in the double-selected schools only complete the survey once, and the data are merged with both datasets for which the school was sampled. No methodology information explicitly specifies which questionnaire is used when a school is double-sampled; however, this explains how a student in the national sample might possibly receive a state/local/tribal version of the questionnaire.

Table A-1 notes the sample years when integrated variables were included only on the national YRBSS.

**Table A-1: Integrated variables only included on the national YRBSS, 1993-2013**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **1993** | **1995** | **1997** | **1999** | **2001** | **2003** | **2005** | **2007** | **2009** | **2011** | **2013** |
| SCHLTALKAIDSNUMYR | x |  |  |  |  |  |  |  |  |  |  |
| CIGAGEREG |  | x |  |  |  |  |  |  |  |  |  |
| MOMEDUC | x | x | x |  |  |  |  |  |  |  |  |
| DADEDUC | x | x | x |  |  |  |  |  |  |  |  |
| CIGEVDLY |  | x | x |  |  |  |  |  |  |  |  |
| CGARDAYMO |  |  | x |  |  |  |  |  |  |  |  |
| SEXPARTAGE |  |  | x |  |  |  |  |  |  |  |  |
| DVLAST |  |  |  | x | x |  |  |  |  |  |  |
| CIGBRAND |  |  |  | x | x | x |  |  |  |  |  |
| DENTLAST |  |  |  | x | x | x |  |  |  |  |  |
| SUNSCRN |  |  |  | x | x | x | x | x | x | x | x |
| ECSTASYFRQEV |  |  |  |  | x |  |  |  |  |  |  |
| HALLFRQEV |  |  |  |  | x |  |  |  |  |  |  |
| EXERINJRYFRQYR |  |  |  |  | x | x | x | x |  |  |  |
| STBELTDRIV |  |  |  |  | x | x |  |  |  |  |  |
| ASTHMAEV |  |  |  |  |  | x |  |  |  |  |  |
| ASTHMAATKYRR |  |  |  |  |  | x |  |  |  |  |  |
| CIGSIDMO |  |  |  |  |  | x | x |  |  |  |  |
| COMPHRSNUMDAY |  |  |  |  |  | x | x |  |  |  |  |
| HALLFRQEV |  |  |  |  |  | x | x | x | x | x | x |
| DISABILITY |  |  |  |  |  |  | x |  |  |  |  |
| ABSENTDAYMO |  |  |  |  |  |  | x |  |  |  |  |
| SUNPROTECT |  |  |  |  |  |  | x | x |  |  |  |
| MCYCHLMTFRQYR |  |  |  |  |  |  | x | x | x |  |  |
| HIVTESTEV |  |  |  |  |  |  | x | x | x | x | x |
| HEALTH |  |  |  |  |  |  |  | x |  |  |  |
| EXERHRDDAYWK |  |  |  |  |  |  |  | x | x |  |  |
| EXERMODDAYWK |  |  |  |  |  |  |  | x | x |  |  |
| PEEXERMIN |  |  |  |  |  |  |  | x | x |  |  |
| SLEEPHR |  |  |  |  |  |  |  | x | x | x | x |
| RXDRUGFRQEV |  |  |  |  |  |  |  |  | x |  |  |
| GRADES |  |  |  |  |  |  |  |  | x |  |  |
| TANFRQYR |  |  |  |  |  |  |  |  | x | x | x |
| TEXTDRIVDAYMO |  |  |  |  |  |  |  |  |  | x |  |
| SCHLSTOLEFRQYR |  |  |  |  |  |  |  |  |  | x |  |
| DRNKMILKFRQWK |  |  |  |  |  |  |  |  |  | x |  |
| EATBRKFSTFRQWK |  |  |  |  |  |  |  |  |  | x |  |
| EXERTONEDAYWK |  |  |  |  |  |  |  |  |  | x | x |

1. Centers for Disease Control and Prevention. Retrieved from <https://www.cdc.gov/healthyyouth/data/yrbs/index.htm> [↑](#footnote-ref-1)
2. Centers for Disease Control and Prevention. Retrieved from <https://www.cdc.gov/mmwr/pdf/rr/rr6201.pdf> [↑](#footnote-ref-2)
3. Brogan, Donna. "Sampling error estimation for survey data." *Household sample surveys in developing and transition countries. New York: United Nations Publication* (2005): 447-90. See Chapter XXI, page 449: <https://unstats.un.org/unsd/HHsurveys/pdf/Chapter_21.pdf>. [↑](#footnote-ref-3)
4. SAS/STAT(R) 9.2 User's Guide, Second Edition. Retrieved from [http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug\_surveymeans\_sect007.htm](http://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm%23statug_surveymeans_sect007.htm) [↑](#footnote-ref-4)
5. Paben, Steven. "Comparison of Variance Estimation Methods for the National Compensation Survey." *Proceedings of the Section on Survey Research Methods, American Statistical Association*. 1999. <https://stats.bls.gov/ore/pdf/st990290.pdf>, page 4 [↑](#footnote-ref-5)
6. Lohr, Sharon. *Sampling: design and analysis*. Nelson Education, 2009, <http://www.math.zju.edu.cn/webpagenew/uploadfiles/attachfiles/201335142847631.pdf>, page 369. [↑](#footnote-ref-6)
7. SAS/STAT(R) 9.2 User's Guide, Second Edition. Retrieved from <https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_surveyfreq_sect014.htm> [↑](#footnote-ref-7)