State and Local Youth Risk Behavior Surveys
Weighting Procedures

This document summarizes the procedures that are applied for weighting data from state and local Youth Risk Behavior Surveys (YRBS). The summary describes, in general, the weighting procedures that are applied in surveys for which the YRBS sampling software, PCSample, is used to select a sample. Weighting procedures for surveys that use other sample designs may differ from those described in this document. Questions regarding weighting procedures should be addressed to the statistician weighting the data for your state or local agency.

Weighting in General

For most YRBS sites, it is impractical and unnecessary to administer the YRBS to every student in the population. PCSample selects representative samples of schools and classes within selected schools. The sample is designed so that every eligible student has an equal chance of selection.

The sample is selected in two steps. In the first step, schools are selected with probability proportional to the enrollment of the school. In the second step, classes are selected within schools with equal probability. The questionnaire is administered to all students in sampled classes in the sampled schools.

The objective of the weighting process is to develop sample weights so that the weighted sample estimates accurately represent the entire student population in the state or city. Nonresponse or poor sampling procedures can result in a sample that is not a representative subset of the population. Unweighted results from these samples may not accurately reflect student behaviors and therefore may be misleading.

Figure 1 shows the steps that are used in Westat's weighting adjustments. Each of these steps is described in more detail in the following sections. The boxes in Figure 1 are numbered to correspond to the section numbers.

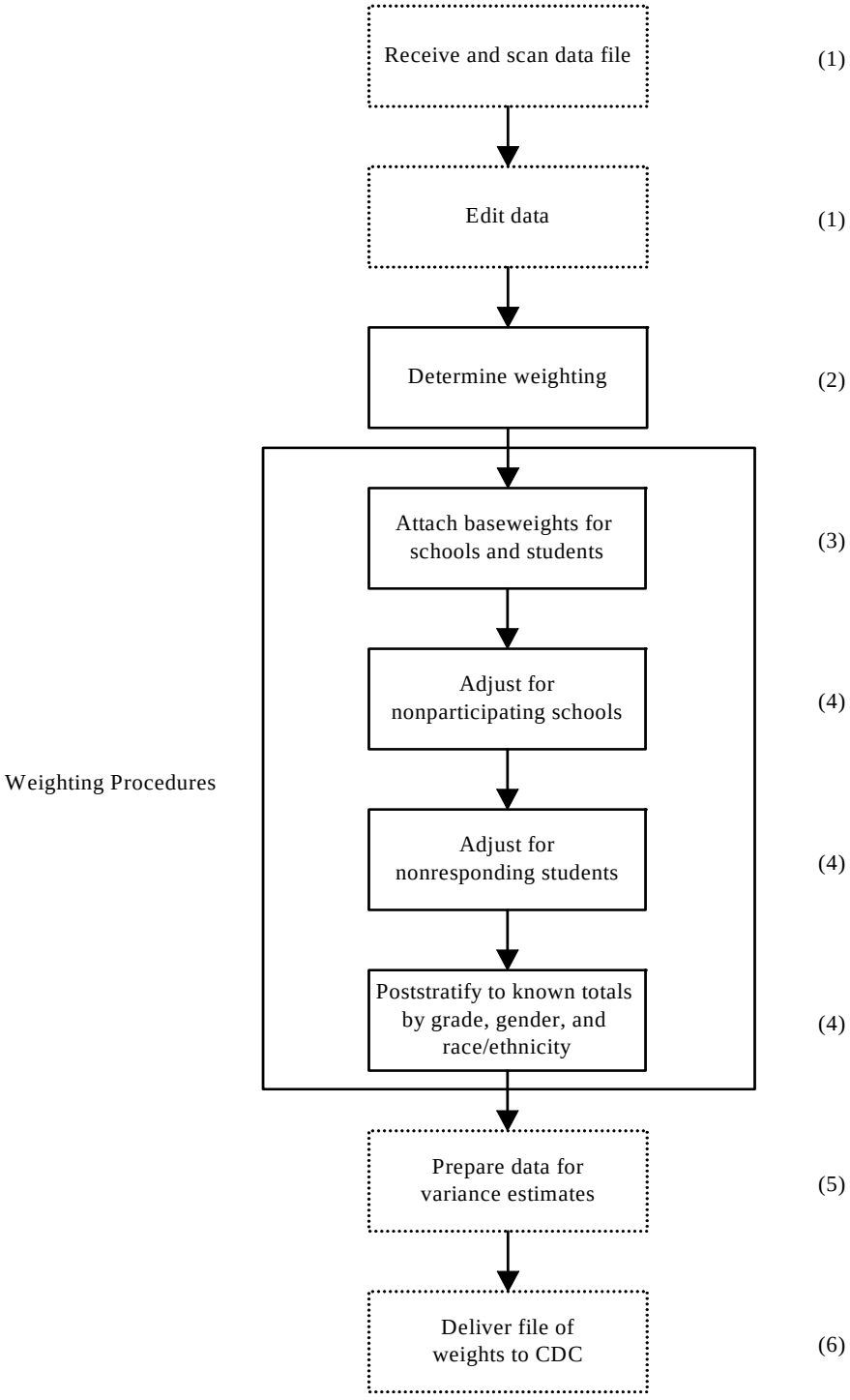

Figure 1. The YRBS Weighting Procedures

1. Prepare the Data

Completed questionnaires from a survey are scanned at Westat, a data file is created, and the file is sent to CDC to be edited. CDC edits the data to identify responses that are inconsistent or otherwise questionable. The edited data are returned to Westat for weighting.

2. Determine if Data can be Weighted

Statisticians at Westat receive the data for a site in the form of initial "unweighted" frequencies. At this point, the statistician determines whether the data can be weighted. To determine if a YRBS data set can be weighted, all of the following conditions must be met:

- Legitimate sampling methods were used (i.e., every student has a known chance of selection and the probabilities of selection can be defined and computed for each sampled student);
- There is enough documentation available to calculate and attach weights (i.e., probabilities of selection can be defined and computed for each sampled student); and
- The overall response rate is at least 60 percent.

The first two conditions are basic requirements for computing the correct probabilities of selection and initial weights. Without this information, weighting is not possible regardless of the response rate. If the sample was selected using PCSample, and if the school and classroom selection procedures were applied properly, and if all work is well documented, these conditions are satisfied. Otherwise, the procedures used to select the sample must be documented completely and carefully.

There are two components to the overall response rate: a school response rate, and a student response rate. Each of these response rates is calculated as follows:

$$\text{School Response Rate} = \frac{\text{Number of Participating Schools}}{\text{Number of Eligible Sampled Schools}}$$

$$\text{Student Response Rate} = \frac{\text{Number of Usable Questionnaires}}{\text{Number of Eligible Students Sampled in Participating Schools}}$$

The overall response rate is calculated as:

$$\textit{Overall Response Rate}^{1} = \left(\frac{\textit{Number of participating schools}}{\textit{Number of eligible sampled schools}}\right) * \left(\frac{\textit{Number of usable questionnaires}}{\textit{Number of eligible students sampled in participating schools}}\right)$$

[1] Rounded to the nearest integer.

The number of usable questionnaires is determined after data have been edited[2]. Only eligible schools and students are counted for determining response rates.

3. Attach Baseweights

PCSample assigns base weights to each sampled student. The weight is equal to the inverse of the probability that the student is selected for the survey. This weight can be thought of as the number of students in the population that are represented by each sampled student.

The weight for each sampled student is computed as follows:

$$Student\ weight = School\ weight * Within\text{-}school\ weight\ .$$

The school weight is based on the probability of selection for the school; and the within-school weight is based on the probability of selection for classes within each sampled school[3]. Samples that are selected by PCSample have the additional property that each sampled student has the same weight, i.e., the sample is "self-weighting."

4. Adjusting the Weights

Adjustments are made to the initial weights to remove bias from the estimates and reduce the variability of the estimate. Westat's standard weighting process for the YRBS involves three adjustments to the weights. Two adjustments are made to account for nonresponse in the sample and one adjustment is made to fine tune the weighted sample estimates to known population characteristics that can affect responses to survey questions. Each of these adjustments is summarized below.

The first adjustment accounts for nonparticipating schools that were sampled. This adjustment is made at the school level and accounts for entire schools that are sampled but are unable, or refuse, to participate. For this adjustment, schools are grouped into three categories based on school enrollment. Within each category, weights of refusing schools are distributed to the participating schools.

The second adjustment is made at the student-level and accounts for eligible students enrolled in sampled classes who fail to complete a questionnaire (e.g., students who are absent on the day the survey is administered, students who do not receive parental permission, students who refuse to participate, or questionnaires that fail the edit and quality control checks). Weights of these nonresponding students in sampled classes are given to responding students in the same class or in classes of similar grade in the same school.

---

[2] When the questionnaire contains more than 54 total questions, a questionnaire is complete if a student answers at least 21 questions and the student answers with b, c, d, e, f, g, or h less than 15 times in a row. When the questionnaire contains 54 or fewer total questions, a questionnaire is complete if a student answers at least 16 questions and the student answers with b, c, d, e, f, g, or h less than 12 times in a row.

[3] Detailed documentation of YRBS sampling procedures is provided in "PCSample Description and Operation." This document is available on request from Westat or CDC.

The final weighting step is to adjust weighted sample totals to known population totals for variables that can affect response to survey questions. Raking ratio estimation, also known as iterative poststratification or raking is used to adjust the weights to two sets of population totals simultaneously. The raking variables used are: (1) grade by gender and (2) race-ethnicity. Weighted sample frequencies in each raking variable are adjusted so that the weighted sample totals of grade by gender and race-ethnicity agree with the true population totals for the state or city.

Additional technical details for these weighting steps are provided in Appendix A to this summary.

5. Attaching Variables for Variance Estimates

Weighted estimates and standard errors are calculated at CDC using SUDAAN. This is a special purpose computer application that calculates estimates and standard errors for data from complex surveys. To use this program, two variables must be defined for calculating standard errors. These variables identify the variance strata and the primary sampling units (PSUs). Variables identifying variance stratum and PSU are created at Westat following weighting.

Values of these variables are based on the procedures that were used to select the sample. In PCSample, schools are selected using implicit stratification that is based on school enrollment. Sampling strata for SUDAAN are defined to consist of either a single certainty school or pairs (or triplets) of noncertainty schools. Pairs (or triplets) of noncertainty schools are grouped according to the order of sample selection. PSU's are comprised of classes within schools for certainty strata and schools within groups for noncertainty strata. More detail regarding the definition of these variables is provided in Appendix A.

6. Final Files

For surveys that are weighted, Westat creates a file for CDC that includes the record ID, the final weights, the variance stratum, and the PSU. The weight file contains all scanned records, including records that CDC subverted due to inconsistent responses and records that Westat deleted due to sampling error. These ineligible records have zero weights, missing variance stratum, and missing PSU on the file.

For surveys that are unweighted, Westat sends a file to CDC containing the record ID and an eligibility variable for identifying the eligible records.

7. Nonstandard Sample Designs

In general, weighting adjustments should be based on procedures used to select the sample. When nonstandard procedures are used to select a YRBS sample, the weighting procedures are tailored to the sample design that was used. Weighting procedures for these surveys are modified as necessary to account for specific sample designs.

Appendix A: Technical Summary of YRBS Weighting

A.1 Initial Weights

Every eligible student is assigned a base weight, which is equal to the inverse of the probability of selection for the student. Student probabilities of selection are calculated from:

$$P(\text{Student is Selected}) = P(\text{School is Selected}) \times P(\text{Class is Selected} \mid \text{School is Selected}) \times P(\text{Student is Selected} \mid \text{School and Class are Selected})$$

For the YRBS, all students in sampled classes are selected so that

$$P(\text{Student is Selected} \mid \text{School and Class are Selected}) = 1.$$

A baseweight is computed for each sampled student as:

$$\begin{aligned} \textit{Student baseweight} &= \frac{1}{P(\textit{Student is Selected})} \\ &= \frac{1}{P(\textit{School is Selected})} \times \frac{1}{P(\textit{Class is Selected} \mid \textit{School is Selected})} \times 1 \\ &= \textit{School baseweight} \times \textit{Within - school baseweight} \end{aligned}$$

Schools are selected with probability proportional to size (PPS), with size defined as school enrollment in the target grades. A baseweight is calculated for each school as:

$$\text{Baseweight for school } i = \begin{cases} \dfrac{\sum \text{Measure of size of all noncertainty schools in the frame}}{n \times \text{measure of size assigned to school } i}, & \text{if school } i \text{ is selected with noncertainty} \\ 1, & \text{if school } i \text{ is selected with certainty} \end{cases}$$

where n is the number of noncertainty schools required in the sample.

The "within-school weight" is equal to the inverse of the conditional probability that the class is selected given the school is selected. PCSample determines this sampling rate so that the resulting probability of selection for each student is equal to the overall sampling rate. Using basic algebra, the required within-school weight can be shown to be equal to:

$$\textit{Within} - \textit{school baseweight for school } i = \frac{1}{f \times \textit{Baseweight for school } i}$$

where $f = \text{the overall sampling rate} = \dfrac{\text{Adjusted student sample size}}{\text{Frame enrollment}}$

and the adjusted student sample size is computed from the number of completes required, adjusted for school nonresponse, student nonresponse, and nonparticipation due to permission requirements.

The resulting overall student probability of selection is then

$$\begin{aligned} P(\text{Student is Selected}) &= P(\text{School is Selected}) \times P(\text{Class is Selected} \mid \text{School is Selected}) \\ &= \frac{1}{\text{Baseweight for school } i} \times \frac{1}{\text{Within} - \text{school baseweight for school } i} \\ &= \frac{1}{\text{Baseweight for school } i} \times f \times (\text{Baseweight for school } i) = f \end{aligned}$$

Thus, each student has the same probability of being selected for the sample, and the resulting sample is "self-weighting."

When there are schools on the frame that have very small enrollments, it is possible that

$$\frac{1}{f \times \text{Baseweight for school } i} > 1.$$

This occurs if the school probability of selection is so small that even if all students in the school are selected, the overall probability for students in the school will be less than the overall sampling rate, *f*. In this case, PCSample increases the measure of size for small schools in such a way that the resulting probabilities of selection will be the same for all eligible students.

A.2 Nonresponse Adjustments

Each eligible student that is sampled represents students in the population, whether or not the eligible sampled student completes a questionnaire. In the weighting adjustments for nonresponse, students in the population that are represented by survey nonrespondents are reassigned to survey respondents. The reassignment attempts to match respondents and nonrespondents with respect to variables that affect response propensity.

Nonresponse adjustment for the YRBS is accomplished with two adjustment steps. The first adjustment accounts for schools that do not participate; and the second adjustment accounts for refusing students in participating schools.

A.3 School Nonresponse Adjustment

To adjust for school nonresponse, each sampled school is assigned to one of three groups based on school enrollment in the target grades: large schools, medium schools, and small schools. The

groups are constructed so that each group has approximately the same total enrollment. Within each group, school-level nonresponse adjustments are calculated as:

$$\textit{School}\ \text{adjustment factor} = \frac{\sum_{\textit{Selected schools}} \left(\textit{School baseweight} \times \textit{School enrollment}\right)}{\sum_{\textit{Participating schools}} \left(\textit{School baseweight} \times \textit{School enrollment}\right)}$$

The adjusted school weight is calculated as:

$$\text{Adjusted school weight for school } i = \begin{cases} \text{Baseweight for school } i \times \text{School adjustment factor, if school } i \text{ participates} \\ 0, \text{ if school } i \text{ refuses} \end{cases}$$

Cells that have low frequencies (less than 3 schools) and cells that have very high adjustment factors (greater than 2.5) may be collapsed with other cells for calculating the final adjustments.

A.4 Student Nonresponse Adjustment

In schools that participate, sampled students may fail to complete a questionnaire for a variety of reasons including absence, refusal to participate, attendance at special functions outside the classroom, or lack of parental permission. Student nonresponse also arises when questionnaires fail the edit and quality control checks. The student-level nonresponse adjustment accounts for loss of sampled students in participating schools.

Adjustment cells for the student-level adjustment are based on classrooms within schools. Cells with low frequencies (less than 15 students) or very high adjustment factors (greater than 2.5) may be collapsed with other cells using criteria that take into account the school size category and the modal grade of the class.

Within each adjustment cell, a student nonresponse adjustment factor is computed from:

$$\text{Student adjustment factor} = \frac{\sum_{\text{Eligible students}} \text{Student weight}}{\sum_{\text{Completed surveys}} \text{Student weight}}$$

where *Student weight = Adjusted school weight x Within-school weight.*

The resulting final adjusted student weights are:

$$\text{Adjusted student weight for student } j = \begin{cases} \text{Student weight for student } j \times \text{Student adjustment factor, if student } j \text{ responds} \\ 0, \text{ if student } j \text{ refuses} \end{cases}$$

A.5 Raking

The final weighting step adjusts the weights so that weighted sample estimates match known marginal population totals by grade and gender and by race-ethnicity. This technique is called raking. Raking is often used when marginal totals are known, but interior cell counts can only be estimated from the sample. The weights are adjusted to the first marginal distribution, or set of control totals, then the second, and so on. This sequence is repeated until the adjusted weights converge to the control totals in each dimension.

For the YRBS, adjustment cells for raking are based on classification of students by grade and gender and by race-ethnicity. The first raking dimension is by grade and gender, consisting of eight adjustment cells of males and females in each of grades 9, 10, 11, and 12. Each responding sampled student is assigned to an adjustment cell based on the grade and gender reported in the questionnaire.

The second raking dimension is by race-ethnicity. Race-ethnicity is grouped into at most three categories based on the race-ethnicity distribution in the population. Each category must contain at least five percent of the race-ethnicity distribution. The remaining students not in these highly concentrated categories are placed in a separate category, “other”. This “other” category also includes students who reported more than one race-ethnicity categories on the questionnaire. Each responding sampled student is assigned to an adjustment cell based on the race-ethnicity reported in the questionnaire.

Control totals for each cell are provided by each state or local agency using school enrollment tabulations. Within each cell, adjustment factors are computed as:

$$Raking\ adjustment\ factor\ =\ \frac{Control\ total}{\sum_{Responding\ students} Adjusted\ student\ weight}$$

The final weight for each responding student is computed as:

$$Final\ weight\ =\ Raking\ adjustment\ factor\ \times\ Adjusted\ student\ weight$$

Sampled students reporting their grade as “Ungraded” or “Other” are not included in the poststratification adjustment. These students retain their weight from the nonresponse adjustment.

Occasionally a completed questionnaire may have missing responses for the items used in raking. For the raking step, missing responses for grade, gender, and race-ethnicity are imputed so that all responding sampled students can be assigned to an appropriate adjustment cell. Hot-deck imputation is used, where students with missing items (recipients) are filled in with reported items from other students (donors). Donors and recipients are grouped into cells that are similar in some auxiliary variables. Within each cell, donors and recipients are matched randomly. Values of these imputed variables are not included in the data file sent to the site.

A.6 Preparation for Variance Estimation

Variances for the YRBS survey data are estimated at CDC using SUDAAN. Estimates of variability for data from complex designs require specialized methods designed specifically for this purpose. SUDAAN is a software package that was developed specifically to compute estimates and variances for data from complex designs.

To use this software, two additional variables are required that identify the sampling stratum (STRATUM) and primary sampling unit (PSU) assignments for participating schools and classes. Although not strictly part of the weighting adjustments, values of these variables depend on the sample design used for the survey.

In PCSample, schools are sorted prior to sampling based on enrollment in the target grades. Very large schools are sampled with certainty. Noncertainty schools are sampled using systematic sampling with probability proportional to enrollment. Sampling strata for SUDAAN consist of either a single certainty school or pairs (or triplets) of noncertainty schools.

Within certainty schools, each class comprises a PSU, so that strata formed from certainty schools can have several PSU's. Certainty schools in which only a single class is sampled are combined into strata with schools of similar size and locale.[4]

Noncertainty schools are grouped into pairs according to the order they are sampled.[5] If there are an odd number of noncertainty schools then the final group is a triplet. Each pair (or triplet) comprises a stratum for the noncertainty schools; and each school comprises a PSU.

[4]See description of method of "collapsed strata" in Cochran, William G., *Sampling Techniques*, Wiley, 1977, pp. 139-140.

[5]See the discussion of estimators for systematic sampling with unequal probabilities in Wolter, Kirk M., *Introduction to Variance Estimation*, Springer-Verlag, 1985, pp. 286-287.