Global School-based Student Health Survey

2013 GSHS Data User's Guide

CDC
Centers for Disease Control and Prevention

World Health Organization

2013 GSHS Data User's Guide

Overview

Introduction

This document provides technical information about Global School-based Student Health Survey (GSHS) data files. It

- provides background information about the GSHS,
- explains how data are edited for quality and consistency,
- explains how analysis variables are calculated,
- explains how data files sent to countries for their surveys differ from public-use data files posted on the WHO and CDC web sites, and
- provides references to further analytic guidance.

It is intended for analysts familiar with statistical software packages and with survey data in general.

Contents

This publication contains the following topics:

GSHS Background

Introduction

The Global School-based Student Health Survey (GSHS) was developed by the World Health Organization (WHO) and the Centers for Disease Control and Prevention (CDC) in collaboration with UNICEF, UNESCO, and UNAIDS.

Goal

The goal of the GSHS is to obtain systematic information from students to support school health and youth health programs and policies globally.

Purpose

The purpose of the GSHS is to provide data on health behaviors and protective factors among students to

- help countries develop priorities, establish programs, and advocate for resources for school health and youth health programs and policies
- allow international agencies, countries, and others to make comparisons across countries regarding the prevalence of health behaviors and protective factors and
- establish trends in the prevalence of health behaviors and protective factors by country for use in evaluating school health and youth health promotion programs.

Methodology

The GSHS is a school-based survey conducted primarily among students aged 13–17 years. The GSHS uses a standardized scientific sample selection process; common school-based methodology; and core questionnaire modules, core-expanded questions, and country-specific questions that are combined to form a questionnaire that can be administered during one regular class period.

Core Questionnaire Modules

The 10 core questionnaire modules address the leading causes of morbidity and mortality among children and adults worldwide. They are

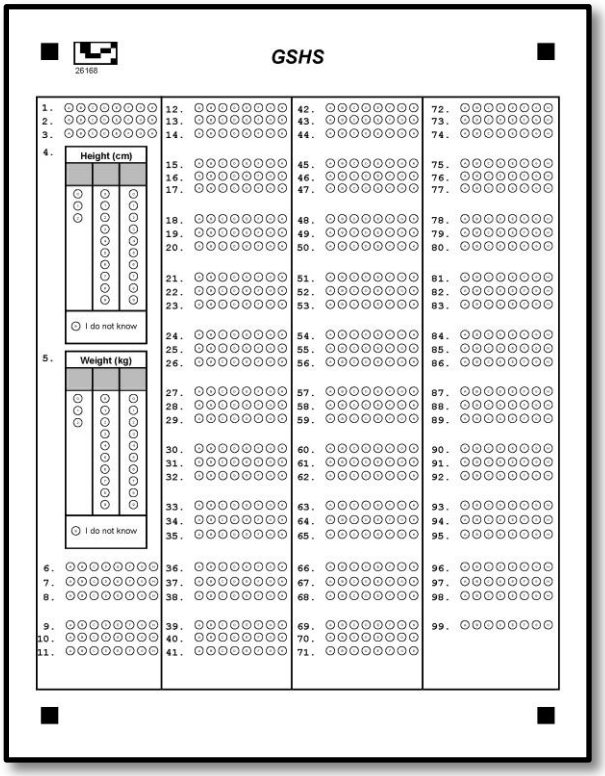
- Alcohol use
- Dietary behaviors
- Drug use
- Hygiene
- Mental health
- Physical activity

- Protective factors
- Sexual behaviors that contribute to HIV infection, other sexually-transmitted infections, and unintended pregnancy
- Tobacco use
- Violence and unintentional injury

Countries must include at least six of the 10 core modules with no changes in their questionnaires. Countries can add questions from a core-expanded question list or from other sources.

Administration

The GSHS is self-administered. Students enter their answers on a computer scannable answer sheet.



Survey administrators and coordinators collate the answer sheets and send them to CDC for scanning and processing. After the answer sheets are scanned, data are processed using the same data edits for each country to ensure comparability across countries.

For More Information

GSHS core questionnaires, core-expanded questions, item rationale, country questionnaires, fact sheets, public-use data files, and other GSHS information can be found on the Global School-based Student Health Survey web sites of WHO and CDC.

Data Edits

Introduction

The following edits are performed on GSHS data:

- out of range edits
- multiple response edits
- logical consistency edits
- height, weight, and body mass index (BMI) edits
- variable-level edits
- record-level edits

The data edits are explained in detail below.

Out of Range Edits

Data are checked for out of range responses.

- If a student selects a response that does not correspond to one of the possible responses for a question, then the question is set to missing.

Example: If "A" and "B" are the valid response options for a question and a student selects "C", "D", "E", "F", "G", or "H," then the question is set to missing for that student.

Multiple Response Edits

Data are checked for multiple responses to a single question.

- If a student selects more than one response for a question, then the question is set to missing.

Example: If "A", "B," and "C" are the valid response options for a question and a student selects both "B" and "C," then the question is set to missing for that student.

Logical Consistency Edits

Data are checked for logical consistency between related questions.

- If a student selects responses from two related questions that conflict logically, then both questions are set to missing unless one of the questions is the question about the age of that student.

- If a student selects a response to a question that conflicts with the age reported by the student then, the question is set to missing but age is left unchanged. This preserves the responses to the age question.

Example: Consider the following three questions.

1. How old are you?
   A. 11 years old or younger
   B. 12 years old
   C. 13 years old
   D. 14 years old
   E. 15 years old
   F. 16 years old
   G. 17 years old
   H. 18 years old or older

6. How old were you when you first tried a cigarette?
   A. I have never smoked cigarettes
   B. 7 years old or younger
   C. 8 or 9 years old
   D. 10 or 11 years old
   E. 12 or 13 years old
   F. 14 or 15 years old
   G. 16 or 17 years old
   H. 18 years old or older

7. During the past 30 days, on how many days did you smoke cigarettes?
   A. 0 days
   B. 1 or 2 days
   C. 3 to 5 days
   D. 6 to 9 days
   E. 10 to 19 days
   F. 20 to 29 days
   G. All 30 days

If a student selects "A. I have never smoked cigarettes" for question 6 and also selects "C. 3 to 5 days" for question 7, then the responses are not logically consistent and both are set to missing for that student.

If a student selects "D. 14 years old" for question 1 and also selects "G. 16 or

17 years old" for question 6, then the responses are not logically consistent and the response to question 6 is set to missing for that student. The response to question 1 is not changed.

The following standard logical consistency edits are applied to questions from the GSHS core questionnaire modules:

Violence and Unintentional Injury

1. Q17=A AND Q18=B,C,D,E,F,G,H
2. Q17=A AND Q19=B,C,D,E,F,G,H
3. Q20=A AND Q21=B,C,D,E,F,G,H

Tobacco Use

4. Q28=A AND Q29=B,C,D,E,F,G
5. Q28=A AND Q31=B,C,D
6. Q1=A AND Q28=E,F,G,H
7. Q1=B AND Q28=F,G,H
8. Q1=C AND Q28=F,G,H
9. Q1=D AND Q28=G,H
10. Q1=E AND Q28=G,H
11. Q1=F AND Q28=H
12. Q1=G AND Q28=H

Alcohol Use

13. Q34=A AND Q35=B,C,D,E,F,G
14. Q34=A AND Q36=C,D,E,F,G
15. Q34=A AND Q37=B,C,D,E,F,G
16. Q34=A AND Q38=B,C,D
17. Q34=A AND Q39=B,C,D
18. Q1=A AND Q34=E,F,G,H
19. Q1=B AND Q34=F,G,H
20. Q1=C AND Q34=F,G,H
21. Q1=D AND Q34=G,H
22. Q1=E AND Q34=G,H
23. Q1=F AND Q34=H
24. Q1=G AND Q34=H

Drug Use

25. Q40=A AND Q41=B,C,D,E
26. Q40=A AND Q42=B,C,D,E
27. Q40=A AND Q43=B,C,D,E
28. Q1=A AND Q40=E,F,G,H
29. Q1=B AND Q40=F,G,H
30. Q1=C AND Q40=F,G,H
31. Q1=D AND Q40=G,H
32. Q1=E AND Q40=G,H

33. Q1=F AND Q40=H
34. Q1=G AND Q40=H

Sexual Behaviors That Contribute to HIV Infection, Other STI, and Unintended Pregnancy

35. Q44=B AND Q45=B,C,D,E,F,G
36. Q44=B AND Q46=B,C,D,E,F,G
37. Q44=B AND Q47=B,C
38. Q44=B AND Q48=B,C,D
39. Q1=A AND Q45=E,F,G,H
40. Q1=B AND Q45=F,G,H
41. Q1=C AND Q45=F,G,H
42. Q1=D AND Q45=G,H
43. Q1=E AND Q45=G,H
44. Q1=F AND Q45=H
45. Q1=G AND Q45=H

In addition, data processing staff write additional logical consistency edits as needed to check core-expanded or country specific questions that have been added to a questionnaire. Core-expanded and country added questions are not edited against core module questions even if they could be considered logically related. This is to ensure that core module questions remain comparable across countries.

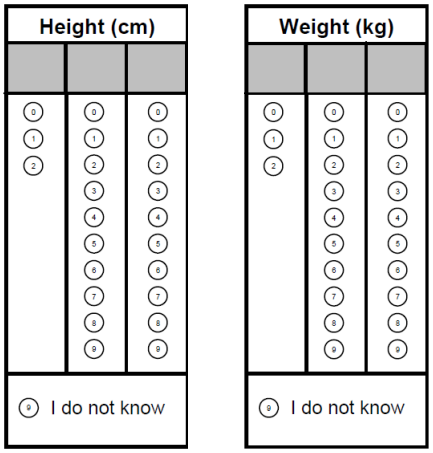
Height, Weight, and Body Mass Index (BMI) Edits

Typically, survey staff weigh and measure the height of each students before survey administration and write the results on slips of paper that are given to each student. Students enter their height and weight onto the GSHS answer sheet during survey administration.

Height and *weight*, along with *sex* and *age*, are used to calculate body mass index (BMI) and then indicators for underweight, overweight, and obesity. WHO Growth Reference Data are used to determine underweight, overweight, and obesity. *Height*, *weight*, and *BMI* are edited to ensure that results are plausible before the indicators are calculated.

Height/Weight Edits

Students enter their height in centimeters and weight in kilograms into grids on the answer sheets.



- If a student leaves the first column in *Height* or *Weight* blank, then a "0" is assumed.

- If a student selects more than one number in any column, then the question is set to missing.

- If a student selects "I do not know," then the question is set to missing.

- If any column is unreadable due to incomplete erasures or other problems, then the question is set to missing.

- *Height* and *weight* must both be usable for either to be retained in the data set. If either is set to missing due to a data problem, then the other is also set to missing.

BMI Edits

- If *Height* and *Weight* are usable then *Height* is converted from centimeters to meters and *BMI* is calculated using the following formula:

$$MIB \quad \frac{ight\ in\ kilo \quad Bms}{ight\ in\ m \ \ tB\ s)\ \$}$$

- If *Height* or *Weight* is missing, then *BMI* is set to missing.

Biologically Implausible Value Edits

If *Height*, *Weight*, and *BMI* are useable, additional edits are applied to ensure results are biologically plausible. The cutoffs for biological plausibility were supplied by the Division of Nutrition, Physical Activity and Obesity, CDC.

- If *Age* or *Sex* is missing, then *Height*, *Weight*, and *BMI* are set to missing because biological plausibility cannot be determined without the age and sex of the student.

- If *Height*, *Weight*, or *BMI* falls outside of the ranges in the following table, then they are all set to missing.

Age	Males	Females
≤ 12	Weight: 20.41-136.08 kg Height: 1.02-1.83 m BMI: 11.5-41.0	Weight: 15.88-136.08 kg Height: 1.02-1.83 m BMI: 11.0-40 .0
13-14	Weight: 27.22-181.44 kg Height: 1.27-1.98 m BMI: 13.0-55.0	Weight: 27.22-181.44 kg Height: 1.27-1.98 m BMI: 13.0-55.0
≥ 15	Weight: 31.75-181.44 kg Height: 1.27-2.11 m BMI: 13.0-55.0	Weight: 27.22-181.44 kg Height: 1.27-1.98 m BMI: 13.0-55.0

Variable-Level Edits

Data are checked to ensure that each question has valid data for at least 60% of all students once all other edits have been completed.

- If less than 60% of students have a valid response for a question, then that question is set to missing for all students.

Record-Level Edits

Data are checked to ensure that each student has at least 20 valid responses once all other edits have been completed. Data are also checked to ensure that there are no cases of too many of the same response in a row.

- If a student record does not have at least 20 valid responses after all edits have been applied, then the entire record for that student is deleted.

- If a record has answers with “B”, “C”, “D”, “E”, “F”, “G,” or “H” 15 or more times in a row, then the entire record for that student is deleted.

Analysis Variables

Introduction

The following kinds of analysis variables are available on GSHS data files:

- core module question variables
- core-expanded and country specific question variables
- dichotomous variables
- supplemental (or calculated) variables
- underweight, overweight, and obese indicator variables
- weight, stratum, and PSU

All variables in a data file will be included in its codebook. Codebooks are specific to individual files and are not interchangeable across data files.

The different types of variables are explained in detail below.

Core Module Question Variables

Core module question variables are named *Q1* through *Q58* in the data file. They contain the edited responses for each core module question.

Example: Consider the following question:

33. Which of your parents or guardians use any form of tobacco?
    A. Neither
    B. My father or male guardian
    C. My mother or female guardian
    D. Both
    E. I do not know

The variable for question 33 is named Q33. *Q33* contains the value 1, 2, 3, 4, or 5 to correspond with the response selected by the student.

Core-expanded and Country Specific Question Variables

Core-expanded question variables and country specific question variables are named *Q59* through *QN99* as needed depending on the number of questions added.

Note: Core-expanded question variables and country specific question variables are not include on public-use data files.

Dichotomous Variables

Each core module question, core-expanded, and country specific question except demographic questions and height and weight have a corresponding dichotomized variable. Dichotomized variables are named *QN6* through *QN99*. Dichotomized variable values divide students into two groups – those

who report a particular behavior or knowledge and those who do not. Dichotomized variables are created by combining responses from the original question into the Response of Interest (ROI) which is the way that variables are most typically reported.

Example: Continuing with the example of question 33 above, *QN33* is the dichotomous variable that corresponds to *Q33*. Q33 has been dichotomized into *QN33* to indicate "students who had parents or guardians who used any form of tobacco" as the ROI. It contains the value 1 for students who selected B, C, or D for *Q33* and the value 2 for students who selected A or E for *Q33*.

Dichotomous variables are created during data processing and are the same for all GSHS data files. Their presence makes it easier to conduct comparable analyses across countries. The original questions are always available, however, for analyses that require different combinations of response options or more detail.

Supplemental Variables

GSHS data files have additional dichotomous variables that are calculated from multiple questionnaire variables or are in addition to other dichotomous variables. Supplemental variable names start with QN followed by a short word or group of letters that helps indicate the content of the variable.

Example: Consider the following two questions:

7. During the past 30 days, how many times per day did you usually eat fruit, such as COUNTRY SPECIFIC EXAMPLES?
   A. I did not eat fruit during the past 30 days
   B. Less than one time per day
   C. 1 time per day
   D. 2 times per day
   E. 3 times per day
   F. 4 times per day
   G. 5 or more times per day

8. During the past 30 days, how many times per day did you usually eat vegetables, such as COUNTRY SPECIFIC EXAMPLES?
   A. I did not eat vegetables during the past 30 days
   B. Less than one time per day
   C. 1 time per day
   D. 2 times per day
   E. 3 times per day
   F. 4 times per day
   G. 5 or more times per day

Q7 and *Q8* have *QN7* and *QN8*, respectively, as their dichotomous variables.

In addition, a supplemental variable is created that combines fruit consumption and vegetable consumption. The supplemental variable *QNFRVGG* contains the value 1 for students whose selections for *Q7* and *Q8* sum to five times per day or more. It contains the value 2 for students whose selections for *Q7* and *Q8* sum to less than five times per day.

Underweight, Overweight, and Obese

Age, sex, and BMI are used to calculate thinness (QNUNWTG), overweight (QNOWTG), and obese (QNOBESEG) indicators. WHO Growth Reference Data are used to determine these indicators for each student with a usable BMI.

- Students are categorized as underweight, i.e., *QNUNWGT* is set to 1, if their BMI is <-2SD from the median for age and sex. If their BMI is greater than -2SD from the median for age and sex, then they are categorized as not underweight, i.e, *QNUNWGT* is set to 2. If their BMI is missing, then *QNUNWGT* is set to missing.

- Students are categorized as overweight, i.e., *QNOWTG* is set to 1, if their BMI is >+1SD from the median for age and sex. If their BMI is less than +1SD from the median for age and sex, then they are categorized as not overweight, i.e., *QNOWTG* is set to 2. If their BMI is missing, then *QNOWTG* is set to missing.

- Students are categorized as obese, i.e., *QNOBESEG* is set to 1, if their BMI is >+2SD from the median for age and sex. If their BMI is less than +2SD from the median for age and sex, then they are categorized as not obese, i.e., *QNOBESEG* is set to 2. If their BMI is missing, then *QNOBESEG* is set to missing.

The BMI-for-age z-score reference tables for boys and for girls from WHO are organized by age in years and months. In the GSHS, students report their age in whole years. Therefore, during processing age in months is approximated by adding six months to the age in years reported by the student. That is, if a student reports that they are 14 years of age, then their age is considered to be 14 years and 6 months. Students who are "11 or younger" are coded as 11 years and 6 months; students who are "18 or older" are coded as 18 years and 6 months.

Note that *QNOWTG* and *QNOBESEG* are not mutually exclusive. That is, if a student is classified as obese, he/she will also be classified as overweight.

Weight, Stratum, and PSU

The weighting process adds *weight*, *stratum*, and *PSU* to every student record in a GSHS data file. All three variables are required to be used when analyzing GSHS data to appropriately represent the weighting process and the 2-stage sample design.

Weight

Weighting allows GSHS results to be generalized to the entire population of students, not just those who took the survey. It allows one student to represent many other students with similar demographic characteristics.

Weighting is necessary for all sample-based surveys since data are not collected from all members of the target population. Weighting accounts for

- the probability of selection of schools and class rooms
- non-responding schools and students, and
- distribution of the population by grade and sex.

For GSHS data to be weighted the following conditions must be met

- the sample from which the data were collected was scientifically selected at both the school and classroom levels.
- all documentation forms were accurately completed.
- a high (>60%) overall response rate was obtained.

The weighting formula used to calculate *weight* for most GSHS data sets is

$$W2 \quad f2 \quad f3B$$

The table below shows what each variable in the formula represents.

Variable		Represents
Base weight	W1	The inverse of the probability of selecting each school
	W2	The inverse of the probability of selecting each class room
Non response adjustment	f1	A school-level non response adjustment factor
	f2	A student-level non response adjustment factor calculated by class room
Post stratification adjustment	f3	A post stratification adjustment factor calculated by sex within grade

Stratum

Stratum describes the 2-stage sample design used for the GSHS. *Stratum* reflects the first level of the GSHS sample selection process, that is, schools.

Stratum is assigned sequentially starting with the school with the largest enrolment of students in the grades/sections/levels/forms that 13-17 year olds usually attend and continuing through the list to the school with the smallest enrolment.

Each stratum usually consists of one or two schools depending on the size of the school. Schools with a very large enrolment that are selected with certainty are assigned their own stratum. The remaining schools are combined in groups of two and each group of two schools shares a stratum.

PSU

PSU also describes the 2-stage sample design used for the GSHS. *PSU* reflects the second level of the GSHS sample selection process, that is, class rooms.

PSU's are assigned sequentially starting with the class rooms in the schools with the largest enrolment of students and continuing through the list to the class rooms in the schools with the smallest enrolment.

Each PSU usually consists of one or more class rooms within a school. Each sampled class room in the very large schools that were selected with certainty is assigned its own PSU. Sampled class rooms in each of the remaining schools share a PSU.

Public-use Data Files

Introduction

Availability of public-use GSHS data files is determined by the Global School-based Student Health Survey (GSHS) Data Release and Publication Policies and Procedures.

Data file availability and data file contents are reviewed briefly below.

Availability

GSHS public-use data files and technical documentation become available for download from the WHO and CDC web sites two years after the country approves the final report.

Data files may be downloaded and used in analyses and publications. No specific permission is required. It is recommended, however, that lead authors of these publications notify CDC of their intent, to help avoid duplication of analytic ideas.

Contents

Public-use GSHS data files contain data only from core module questions. Data from core-expanded questions or country added questions are not included in the public-use data files. School identifiers are not included in the public-use datasets.

Complete country questionnaires are posted on the WHO and CDC web sites, however, so you can see all of the questions that were asked. If you want to analyze results from core-expanded or other questions added to a country questionnaire, please see the Country Contact List for information on contacting the country representative and discussing your request to use their full data files.

More Information on Data Analysis

GSHS surveys employ a two-stage sampling design. Therefore, to analyze GSHS data correctly, statistical software packages that account for this sampling design must be used.

The Youth Risk Behavior Survey (YRBS), conducted by the CDC, uses a comparable sampling design. Software for Analysis of YRBS Data provides guidance on using SUDAAN, SAS, STATA, SPSS, and Epi Info for the analysis of YRBS data. Because GSHS and YRBS methods are comparable, this paper is a good resource for guidance on analyzing GSHS data as well.