Global School-Based Student Health Survey (GSHS)

Data Processing and Weighting

World Health Organization

Overview

- GSHS cleaning, weighting and analysis* have been standardized to ensure the integrity of the global system and comparability of results.
- WHO has developed a set of R code to automate these tasks.
- These slides present how GSHS data is cleaned and weighted and the input files needed to use the R code.

World Health Organization

* Standard descriptive analysis – this is by no means an exhaustive analysis. Countries are encouraged to perform additional analyses.

Data Cleaning

World Health Organization

Overview

- As part of the standard cleaning process, GSHS data are checked for quality and consistency.
- Checking occurs at both the variable level as well as at the record level.
  - Thus, it is possible for a **variable** to be dropped or a student's **entire response** to be dropped.
- The following slides in this section describe the data edits performed.

Out-of-range edits

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

World Health Organization

Multi-response edits

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

- GSHS questions never allow for multiple responses.

World Health Organization

BMI-related edits

- BMI is calculated using the height and weight measures reported in each student's response.
  - If either height or weight are missing, BMI is set to missing.

- Height, weight and BMI are checked to see if they are outside the biologically plausible range*.
  - If height, weight or BMI are implausible, all are set to missing.
  - If age or sex is missing, height, weight and BMI are set to missing since plausible ranges vary by age and sex and it would be impossible to determine plausibility.

World Health Organization

* With a z-score <-5 or >5 according to WHO Growth Standards.

Logical consistency edits

- Logical consistency checks are made for questions in 6 of the core modules.
- These checks ensure that responses are *internally consistent*
  - Example: If a student responds that they did not clean their teeth in the past 30 days but in the following question responds that the toothpaste usually used to brush their teeth in the past 30 days contains fluoride – these responses would not be internally consistent.
- If a check fails, then the responses to **both** questions are set to missing *except* if one of the questions is AGE (age is never set to missing).
- Consistency checks are not exhaustive and there are no consistency checks done for core-expanded or country-specific questions.
- All 46 edits are listed on the following slide.

World Health Organization

Hygiene

1. HY_CLTEETH = A AND HY_FLUORIDE = B,C,D

Injury

2. IN_TIMESINJ = A AND IN_TYPEINJ = B, C, D, E, F, G, H
3. IN_TIMESINJ = A AND IN_CAUSEINJ = B, C, D, E, F, G, H

Tobacco Use

4. TO_TRIEDCIG = B AND TO_AGECIG = B, C, D, E, F, G, H
5. TO_TRIEDCIG = B AND TO_DAYSCIG = B, C, D, E, F, G
6. DE_AGE = A AND TO_AGECIG = E,F,G,H
7. DE_AGE = B AND TO_AGECIG = F,G,H
8. DE_AGE = C AND TO_AGECIG = F,G,H
9. DE_AGE = D AND TO_AGECIG = G,H
10. DE_AGE = E AND TO_AGECIG = G,H
11. DE_AGE = F AND TO_AGECIG = H
12. DE_AGE = G AND TO_AGECIG = H

Alcohol Use

13. AL_AGE = A AND AL_DAYS = B, C, D, E, F, G
14. AL_AGE = A AND AL_DRINKS = B, C, D, E, F, G
15. AL_AGE = A AND AL_INAROW = B, C, D, E, F, G, H
16. AL_AGE = A AND AL_SOURCE = B, C, D, E, F
17. AL_AGE = A AND AL_TROUBLE = B, C, D, E, F
18. AL_AGE =A AND AL_DRUNK = B, C, D, E, F
19. DE_AGE = A AND AL_AGE = E,F,G,H
20. DE_AGE = B AND AL_AGE = F,G,H
21. DE_AGE = C AND AL_AGE = F,G,H
22. DE_AGE = D AND AL_AGE = G,H
23. DE_AGE = E AND AL_AGE = G,H
24. DE_AGE = F AND AL_AGE = H
25. DE_AGE = G AND AL_AGE = H

Drug Use

26. DR_AGE = A AND DR_CANLIFE = B, C, D, E, F
27. DR_AGE = A AND DR_CAN30 = B, C, D, E, F
28. DR_AGE = A AND DR_AMPHLIFE = B, C, D, E, F
29. DE_AGE = A AND DR_AGE = E,F,G,H
30. DE_AGE = B AND DR_AGE = F,G,H
31. DE_AGE = C AND DR_AGE = F,G,H
32. DE_AGE = D AND DR_AGE = G,H
33. DE_AGE = E AND DR_AGE = G,H
34. DE_AGE = F AND DR_AGE = H
35. DE_AGE = G AND DR_AGE = H

Sexual Behaviors

36. DE_AGE = A AND SX_AGE = E,F,G,H
37. DE_AGE = B AND SX_AGE = F,G,H
38. DE_AGE = C AND SX_AGE = F,G,H
39. DE_AGE = D AND SX_AGE = G,H
40. DE_AGE = E AND SX_AGE = G,H
41. DE_AGE = F AND SX_AGE = H
42. DE_AGE = G AND SX_AGE = H
43. SX_EVERSEX = B AND SX_AGE = B, C, D, E, F, G, H
44. SX_EVERSEX = B AND SX_NUMBER = B,C,D,E,F,G
45. SX_EVERSEX = B AND SX_CONDOM = B,C
46. SX_EVERSEX = B AND SX_BC = B,C,D,E,F,G,H

World Health Organization

Variable-level edits

- After all other checks have been implemented, each variable is checked to ensure at least 60% of students have responded.
- If the response rate for a variable is less than 60%, the variable is set to missing for all students.

Record-level edits

- After all other checks have been implemented, each record is checked to ensure there are at least 20 valid responses
- If a student's response has fewer than 20 valid responses, the response is deleted.
- If a record has 15 or more identical responses in a row, other than "A", the entire record is deleted.

World Health Organization

Weighting

Overview

- Once data have been cleaned, the weighting process can begin.
- Weighting accounts for:
  - the probability of selection of schools and classes
  - non-responding schools, classes and students
  - the distribution of the target population (i.e. students in the targeted grades) by grade and sex
- In addition to analysis weights, PSU and Stratum will also be generated which inform the statistical software about the design of your sample.

World Health Organization

Requirements

- All of the following conditions must be met in order to weight GSHS data:
  - the sample was scientifically selected from an up-to-date and complete sampling frame
  - all school-level and class-level forms were accurately completed
  - a high (>60%) overall response rate was obtained

World Health Organization

Weight calculation

- The formula used to calculate analysis weights for most GSHS data sets is

weight = w1 * w2 * f1 * f2 * f3

where:

Base weight
- **w1** = the inverse probability of selecting each school
- **w2** = the inverse probability of selecting each class

Non-response adjustment
- **f1** = a school-level non-response adjustment factor
- **f2** =a student-level non-response adjustment factor (calculated per class)

Post-stratification adjustment
- **f3** = a post-stratification adjustment factor (calculated by sex within each grade)

World Health Organization

PSU and Stratum

- PSU and Stratum describe the complex sample design of the survey
- These numbers are generated as follows:
  - **Schools selected with certainty*** : assign a unique stratum to each school and a unique PSU to each class in each school
  - **All other schools :** sort schools by school weight** and group schools into pairs (if there is an odd number, make one group of three), assign a unique stratum to each pair of schools (or group of three) and a unique PSU to all classes within a given school

World Health Organization

* Large schools with a school weight of 1.0.
** w1 in the weight calculation

PSU and Stratum - example

School Weight	School	Classes	Stratum	PSU
1.0	A	1	1	1
		3	1	2
		6	1	3
1.0	B	2	2	4
		4	2	5
		6	2	6
1.0	C	1	3	7
		3	3	8
		4	3	9
		6	3	10
1.27	E	1	4	11
		2	4	11
		3	4	11
1.38	F	1	4	12
		3	4	12
		5	4	12
1.79	G	2	5	13
		4	5	13
		6	5	13
		8	5	13
1.83	H	1	5	14
		3	5	14
		5	5	14
1.90	I	3	5	15
		6	5	15
		9	5	15

Schools selected with certainty

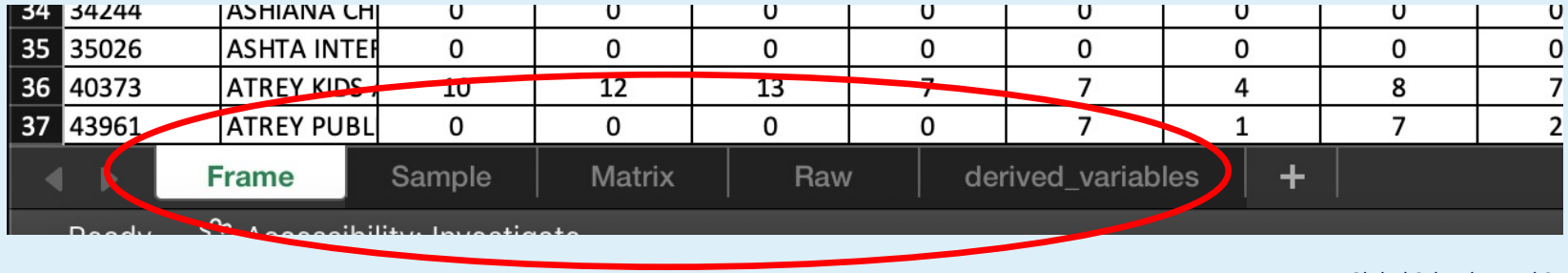
All other schools (i.e. smaller schools)

World Health Organization

Input Files

Overview

- In order to use the standardized code for cleaning, weighting and analyzing your GSHS data, one Excel file with 5 sheets must be prepared:
  1. the raw dataset – named 'Raw'
  2. the sampling frame – named 'Frame'
  3. the sample – named 'Sample'
  4. the analysis matrix – named 'Matrix'
  5. The variables to be derived together with logical conditions named 'derived_variables'
- These files together contain all the necessary information to perform the cleaning and weighting and produce the standard descriptive output.
- The naming, content and structure of these files must be as described in these slides. WHO colleagues can assist in constructing these files correctly and/or verify these files are correctly constructed prior to running the R scripts.



Raw Dataset (named ‘Raw’)

- The raw dataset file is a typical Excel dataset file with one row per student, one column per variable.
- Question responses should be A, B, C, etc. Do not recode these responses to numbers.
- ID variables for schools and classes must be named **school_id** and **class_id**
- Height must be in cms and weight must be in kgs. These variables must be named **height** and **weight**.

World Health Organization

Raw Dataset (raw_data.xlsx)

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	
1	school_id	class_id	Page	q1	q2	q3	height	weight	q6	q7	q8	q9	q10	q11	q12	
2	1	4	2	E	A	B	166	55		C	B	C	B	B	B	C
3	1	4	3	E	A	B	165	50	C	C	B	G	F	B	A	E
4	1	4	4	F	B	B	158	60	C	G	B	A	D	D	A	E
5	1	4	5	F	A	B	167	56	C	C	C	D	B	B	B	C
6	1	4	6	D	A	B	161	46	C	C	B	B	B	B	C	E
7	1	4	7	E	A	B	168	60	C	B	D	D	D	D	G	C
8	1	4	8	F	A	B	169	60	C	B	B	C	E	B	B	C
9	1	4	9	F	B	B	164	72	C	D	D	E	B	C	A	D
10	1	4	10	F	B		152	55	B	A	A	C	B	A	B	C
11	1	4	11	F	B	B	153	55	C	C	G	B	B	B	B	D
12	1	4	12	F	B	B	156	50	A	B	D	D	E	A	C	D
13	1	4	13	E	A	B	170	56	E	B	B	C	E	D	B	E
14	1	4	14	D	A	B	160	54	C	G	E	D	D	D	A	C

World Health Organization

Sampling frame (‘Frame’)

- The sampling frame is used for the post-stratification adjustment only.
- It contains the number of students enrolled in each school in the original sample frame by **grade** and **sex**.
- A **category** variable must be present indicating which stratum each school is in (if the sample was implicitly or explicitly stratified) else this variable can place all schools in a single stratum (e.g. all schools have type = “national”).
- There should be **2 columns** (one for each sex) **per grade**
- The number and ordering of grades **must match the question** students were asked about which grade they were in.

World Health Organization

Sampling frame (‘Frame’)

3. In what grade/class/standard are you?

   A. Form 1
   B. Form 2
   C. Form 3
   D. Form 4
   E. Form 5

Response options A-E appeared in the questionnaire.

The frame contains **2 columns per response** option containing the number of **boys** and the number of **girls** enrolled in that grade in that school.

name	A__BOYS	B__GIRLS	B__BOYS	B__GIRLS	C__BOYS	C__GIRLS	D__BOYS	D__GIRLS	E__BOYS	E__GIRLS	category
A							18	30	18	32	national
B							46	42	44	47	national
C							41	36	28	39	national
D							66	52	66	86	national
E	25	20	20	14	37	52					national
F			176	139	165	150					national
G			204	221	173	167					national
H			81	112	84	133					national
I			35	29	14	26					national
J	11	13	9	7	63	70					national
K	12	8	11	11	31	47					national
L	48	93	42	56	82	109					national

World Health Organization

Sample frame ('Frame')

In this example, the sample was implicitly stratified by type of school. The **category** variable thus reflects this information.

name	A__BOYS	A__GIRLS	B__BOYS	B__GIRLS	C__BOYS	C__GIRLS	D__BOYS	D__GIRLS	E__BOYS	E__GIRLS	category
A	23	22	21	19	26	23	0	0	0	0	Both
B	19	28	25	27	18	26	0	0	0	0	Primary
C	12	10	7	9	11	6	21	9	11	15	Primary
D	6	10	10	9	11	6	0	0	0	0	Secondary
E	18	15	5	12	6	10	0	0	0	0	Secondary
F	22	12	10	8	10	18	17	12	12	14	Primary
G	0	0	22	17	22	27	50	42	46	31	Both
H	20	9	19	14	25	12	0	0	0	0	Primary
I	31	25	32	37	21	15	13	17	20	10	Primary
J	11	15	2	20	7	9	0	0	0	0	Primary
K	46	46	28	41	33	32	35	35	36	41	Both
L	18	17	9	15	16	11	0	0	0	0	Both

Important: Be sure all values of **category** are spelled correctly – misspellings would be interpreted as different strata (e.g. "secondary" would be interpreted as a different stratum than "secondery")

World Health Organization

Sample ('Sample')

- The sample file is used to calculate both the base weights and the non-response adjustments.
- It contains one row per **selected** school which is comprised of the following:
  - School ID
  - School weight (**w1** in the weight calculation)
  - Class sampling interval (**w2** in the weight calculation)
  - School participation flag variable (1 or 0)
  - Total number of eligible classes in the school
  - Total number of classes selected in the school
  - For each selected class: class ID, total enrollment and number of participating students
  - Sampling stratum of the school

World Health Organization

Sample

Variables **school_ID**, **SCWWGT** and **SCINTV** contain the school ID, school weight and class sampling interval.

SCHOOL_ID	SCWGT	SCINTV	school_part	TOTCLASS	SELCLASS	CLASS1	CENROL1	STPART1	CLASS2	CENROL2	STPART2	CLASS3	CENROL3	STPART3	Category
1	4.610865177	1.722059972	1	12	3	4	39	34	7	64	46	11	37	35	Both
2	6.177567754	1.285325662	1	10	3	1	52	43	4	39	34	7	53	40	Both
3	7.181422514	1.105656483	1	12	3	2	41	12	4	34	9	6	34	7	Both
4	7.94018636	1	1	6	3	1	40	38	3	33	25	5	26	19	Primary
5	7.94018636	1	1	5	2	2	41	32	4	32	22	0	0	0	Primary
6	7.94018636	1	1	5	3	1	52	35	3	44	36	5	26	22	Primary
7	7.94018636	1	1	3	1	2	65	46	0	0	0	0	0	0	Both
8	7.94018636	1	1	5	3	1	32	27	3	21	19	5	16	11	Both
9	7.94018636	1	1	3	1	2	35	29	0	0	0	0	0	0	Both
10	7.94018636	1	1	5	3	1	22	22	3	18	18	5	7	7	Both
11	7.94018636	1	1	3	1	2	30	27	0	0	0	0	0	0	Secondary
12	7.94018636	1	1	3	2	1	35	26	3	23	19	0	0	0	Secondary
13	7.94018636	1	1	3	1	2	23	23	0	0	0	0	0	0	Both
14	7.94018636	1	1	4	2	1	23	23	3	16	15	0	0	0	Both
15	7.94018636	1	1	3	1	2	9	7	0	0	0	0	0	0	Both

World Health Organization

Sample

Variables **school_part**, **TOTCLASS** and **SELCLASS** contain the school participation flag, the total number of eligible classes and the number of classes selected.

SCHOOL_ID	SCWGT	SCINTV	school_part	TOTCLASS	SELCLASS	CLASS1	CENROL1	STPART1	CLASS2	CENROL2	STPART2	CLASS3	CENROL3	STPART3	Category
1	4.610865177	1.722059972	1	12	3	4	39	34	7	64	46	11	37	35	Both
2	6.177567754	1.285325662	1	10	3	1	52	43	4	39	34	7	53	40	Both
3	7.181422514	1.105656483	1	12	3	2	41	12	4	34	9	6	34	7	Both
4	7.94018636	1	1	6	3	1	40	38	3	33	25	5	26	19	Primary
5	7.94018636	1	1	5	2	2	41	32	4	32	22	0	0	0	Primary
6	7.94018636	1	1	5	3	1	52	35	3	44	36	5	26	22	Primary
7	7.94018636	1	1	3	1	2	65	46	0	0	0	0	0	0	Both
8	7.94018636	1	1	5	3	1	32	27	3	21	19	5	16	11	Both
9	7.94018636	1	1	3	1	2	35	29	0	0	0	0	0	0	Both
10	7.94018636	1	1	5	3	1	22	22	3	18	18	5	7	7	Both
11	7.94018636	1	1	3	1	2	30	27	0	0	0	0	0	0	Secondary
12	7.94018636	1	1	3	2	1	35	26	3	23	19	0	0	0	Secondary
13	7.94018636	1	1	3	1	2	23	23	0	0	0	0	0	0	Both
14	7.94018636	1	1	4	2	1	23	23	3	16	15	0	0	0	Both
15	7.94018636	1	1	3	1	2	9	7	0	0	0	0	0	0	Both

World Health Organization

Sample

Variables **CLASS#**, **CENROL#** and **STPART#** contain the class ID, total enrollment and number of participating students for each class. These 3 columns can be repeated as many times as needed – 1 set of 3 columns per class.

SCHOOL_ID	SCWGT	SCINTV	school_part	TOTCLASS	SELCLASS	CLASS1	CENROL1	STPART1	CLASS2	CENROL2	STPART2	CLASS3	CENROL3	STPART3	Category
1	4.610865177	1.722059972	1	12	3	4	39	34	7	64	46	11	37	35	Both
2	6.177567754	1.285325662	1	10	3	1	52	43	4	39	34	7	53	40	Both
3	7.181422514	1.105656483	1	12	3	2	41	12	4	34	9	6	34	7	Both
4	7.94018636	1	1	6	3	1	40	38	3	33	25	5	26	19	Primary
5	7.94018636	1	1	5	2	2	41	32	4	32	22	0	0	0	Primary
6	7.94018636	1	1	5	3	1	52	35	3	44	36	5	26	22	Primary
7	7.94018636	1	1	3	1	2	65	46	0	0	0	0	0	0	Both
8	7.94018636	1	1	5	3	1	32	27	3	21	19	5	16	11	Both
9	7.94018636	1	1	3	1	2	35	29	0	0	0	0	0	0	Both
10	7.94018636	1	1	5	3	1	22	22	3	18	18	5	7	7	Both
11	7.94018636	1	1	3	1	2	30	27	0	0	0	0	0	0	Secondary
12	7.94018636	1	1	3	2	1	35	26	3	23	19	0	0	0	Secondary
13	7.94018636	1	1	3	1	2	23	23	0	0	0	0	0	0	Both
14	7.94018636	1	1	4	2	1	23	23	3	16	15	0	0	0	Both
15	7.94018636	1	1	3	1	2	9	7	0	0	0	0	0	0	Both

<u>Important</u>: If a class does not participate, enter the class ID and total enrollment and enter 0 for the number of participating students.

Sample

Finally, the **Category** variable contains the sampling stratum of the school. If no explicit or implicit stratification was done, all schools would get the same value (e.g. "national").

SCHOOL_ID	SCWGT	SCINTV	school_part	TOTCLASS	SELCLASS	CLASS1	CENROL1	STPART1	CLASS2	CENROL2	STPART2	CLASS3	CENROL3	STPART3	Category
1	4.610865177	1.722059972	1	12	3	4	39	34	7	64	46	11	37	35	Both
2	6.177567754	1.285325662	1	10	3	1	52	43	4	39	34	7	53	40	Both
3	7.181422514	1.105656483	1	12	3	2	41	12	4	34	9	6	34	7	Both
4	7.94018636	1	1	6	3	1	40	38	3	33	25	5	26	19	Primary
5	7.94018636	1	1	5	2	2	41	32	4	32	22	0	0	0	Primary
6	7.94018636	1	1	5	3	1	52	35	3	44	36	5	26	22	Primary
7	7.94018636	1	1	3	1	2	65	46	0	0	0	0	0	0	Both
8	7.94018636	1	1	5	3	1	32	27	3	21	19	5	16	11	Both
9	7.94018636	1	1	3	1	2	35	29	0	0	0	0	0	0	Both
10	7.94018636	1	1	5	3	1	22	22	3	18	18	5	7	7	Both
11	7.94018636	1	1	3	1	2	30	27	0	0	0	0	0	0	Secondary
12	7.94018636	1	1	3	2	1	35	26	3	23	19	0	0	0	Secondary
13	7.94018636	1	1	3	1	2	23	23	0	0	0	0	0	0	Both
14	7.94018636	1	1	4	2	1	23	23	3	16	15	0	0	0	Both
15	7.94018636	1	1	3	1	2	9	7	0	0	0	0	0	0	Both

World Health Organization

Important: The values of **category** should be the same as the values of **category** in the sampling frame.

Analysis matrix (mapping_matrix.xlsx)

- The analysis matrix contains information about every question in the questionnaire, including:
  - The original question number used in the questionnaire
  - The original question text and response options
  - The standard variable name assigned to this question
  - Information on the indicator(s) to be calculated from the question

- Each row of the matrix corresponds to one indicator, thus if a question has multiple indicators, the question information is repeated.

World Health Organization

Analysis matrix (‘Matrix’)

The **site** column contains the question number from the fielded questionnaire. The values should match the variable labels in the raw dataset.

The **survey_question** and **var_levels** columns contain the standard variable name for the question, the question text and the response options.

	A	B	C	D	E	F	G	H	I
	bin_standard	site	numerator	denominator_resp_reduced	indicator_description	survey_question	var_levels	factsheet_section	factsheet_subtitle
2	DE_AGE	q1				DE_AGE: How old ar	A:11 years old or younger;B:12 years old;C:13 years old;D:14 ye		
3	DE_SEX	q2				DE_SEX: What is yo	A:Male;B:Female		
4	DE_GRADE	q3				DE_GRADE:In what	A:Class 7;B:Class 8;C:Class 9;D:Class 10;E:Class 11;F:Class 12		
5	DB_HEIGHT	height				DB_HEIGHT: How tall are you without your shoes on (in cm)?			
6	DB_WEIGHT	weight				DB_WEIGHT: How much do you weigh without your shoes on?			
7	DB_UNDERWT				DB_UNDERWT: Percentage of students who were underweight (<-2			dietary	Dietary Behaviours
8	DB_OVERWT				DB_OVERWT: Percentage of students who were overweight (>+1SI			dietary	Dietary Behaviours
9	DB_OBESE				DB_OBESE: Percentage of students who were obese (>+2SD from n			dietary	Dietary Behaviours
10	DB_B_FRUITNONE	q6	c('A')		DB_B_FRUITNONE: Percer	DB_FRUIT: During th	A:I did not eat fruit during the past 7 days;B:1 to 3 times during		
11	DB_B_FRUITLESS	q6	c('A','B','C')		DB_B_FRUITLESS: Percent	DB_FRUIT: During th	A:I did not eat fruit during the past 7 days;B:1 to 3 times during		
12	DB_B_FRUIT1	q6	c('D','E','F','G')		DB_B_FRUIT1: Percentage	DB_FRUIT: During th	A:I did not eat fruit during the past 7 days;B:1 to 3 times during		
13	DB_B_FRUIT2	q6	c('E','F','G')		DB_B_FRUIT2: Percentage	DB_FRUIT: During th	A:I did not eat fruit during the past 7 days;B:1 to 3 times during		
14	DB_B_FRUIT3	q6	c('F','G')		DB_B_FRUIT3: Percentage	DB_FRUIT: During th	A:I did not eat fruit during the past 7 days;B:1 to 3 times during		
15	DB_B_VEGNONE	q7	c('A')		DB_B_VEGNONE: Percent	DB_VEG: During the	A:I did not eat vegetables during the past 7 days;B:1 to 3 times		
16	DB_B_VEGLESS	q7	c('A','B','C')		DB_B_VEGLESS : Percenta	DB_VEG: During the	A:I did not eat vegetables during the past 7 days;B:1 to 3 times		
17	DB_B_VEG1	q7	c('D','E','F','G')		DB_B_VEG1: Percentage c	DB_VEG: During the	A:I did not eat vegetables during the past 7 days;B:1 to 3 times		

World Health Organization

Analysis matrix (‘Matrix’)

The **bin_standard** column contain the name of the indicator.

The **indicator_description** column contains the name and text of the indicator derived from the question.

A	B	C	D	E	F	G	H	I
bin_standard	**site**	**numerator**	**denominator_resp_reduced**	**indicator_description**	**survey_question**	**var_levels**	**factsheet_section**	**factsheet_subtitle**
TO_B_DAYSCIG	q37	c('B','C','D','E','F','G')		TO_B_DAYSCIG: Percentage of students who currently smoked cigarettes (on at least 1 day during the 30 days before the survey)	TO_DAYSCIG: 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	tobacco	Tobacco Use
TO_B_STOPCIG	q38	c('B')	c('A')	TO_B_STOPCIG: Percentage of students who tried to stop smoking cigarettes (among students who smoked cigarettes during the 12 months before the survey)	TO_STOPCIG: During the past 12 months, did you try to stop smoking cigarettes?	A:I did not smoke cigarettes during the past 12 months;B:Yes;C:No		

World Health Organization

Analysis matrix (‘Matrix’)

The **numerator** column contains a list of response options which comprise the numerator of the indicator.

The **denominator_resp_reduced** column contains a list of response options to be excluded from the denominator.

	A	B	C	D	E	F	G	H	I
	bin_standard	**site**	**numerator**	**denominator_resp_reduced**	**indicator_description**	**survey_question**	**var_levels**	**factsheet_section**	**factsheet_subtitle**
5	TO_B_DAYSCIG	q37	c('B','C','D','E','F','G')		TO_B_DAYSCIG: Percentage of students who currently smoked cigarettes (on at least 1 day during the 30 days before the survey)	TO_DAYSCIG: 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	tobacco	Tobacco Use
6	TO_B_STOPCIG	q38	c('B')	c('A')	TO_B_STOPCIG: Percentage of students who tried to stop smoking cigarettes (among students who smoked cigarettes during the 12 months before the survey)	TO_STOPCIG: During the past 12 months, did you try to stop smoking cigarettes?	A:I did not smoke cigarettes during the past 12 months;B:Yes;C:No		

Analysis matrix (‘Matrix’)

The **factsheet_section** and **factsheet_subtitle** columns are used when generating the fact sheet and indicate if the indicator is included in the fact sheet and in which section.

	A	B	C	D	E	F	G	H	I
	bin_standard	**site**	**numerator**	**denominator_resp_reduced**	**indicator_description**	**survey_question**	**var_levels**	**factsheet_section**	**factsheet_subtitle**
5	TO_B_DAYSCIG	q37	c('B','C','D','E','F','G')		TO_B_DAYSCIG: Percentage of students who currently smoked cigarettes (on at least 1 day during the 30 days before the survey)	TO_DAYSCIG: 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	tobacco	Tobacco Use
6	TO_B_STOPCIG	q38	c('B')	c('A')	TO_B_STOPCIG: Percentage of students who tried to stop smoking cigarettes (among students who smoked cigarettes during the 12 months before the survey)	TO_STOPCIG: During the past 12 months, did you try to stop smoking cigarettes?	A:I did not smoke cigarettes during the past 12 months;B:Yes;C:No		

World Health Organization

Derived variables with logical expressions ('derived_variables')

sec_vars	req_vars	log_cond_num	log_cond_denom
DB_B_ALLSSBNONE	DB_SODA,DB_SSB	data$DB_SODA == 'A' & data$DB_SSB == 'A'	All
DB_B_ALLSSBLESS	DB_SODA,DB_SSB	(data$DB_SODA == 'A'\|data$DB_SODA == 'B'|data$DB_SODA == 'C') & (data$DB_SSB == 'A'|data$DB_SSB == 'B'\|data$DB_SSB == 'C')	All
DB_B_ALLSSB2	DB_SODA,DB_SSB	(data$DB_SODA == 'E'\|data$DB_SODA == 'F'|data$DB_SODA == 'G') \| (data$DB_SSB == 'E'|data$DB_SSB == 'F'\|data$DB_SSB == 'G')	All
DB_B_ALLSSB3	DB_SODA,DB_SSB	(data$DB_SODA == 'F'\|data$DB_SODA == 'G') | (data$DB_SSB == 'F'\|data$DB_SSB == 'G')	All
AL_SMOKE_DRINK	AL_DAYS,TO_DAYSCIG	(data$AL_DAYS=='B'\|data$AL_DAYS=='C'|data$AL_DAYS=='D'\|data$AL_DAYS=='E'|data$AL_DAYS=='F'\|data$AL_DAYS=='G')	(data$TO_DAYSCIG=='B'\|data$TO_DAYSCIG=='C'|data$TO_DAYSCIG=='D'\|data$TO_DAYSCIG=='E'|data$TO_DAYSCIG=='F'\|data$TO_DAYSCIG=='G')

Secondary variables to be derived

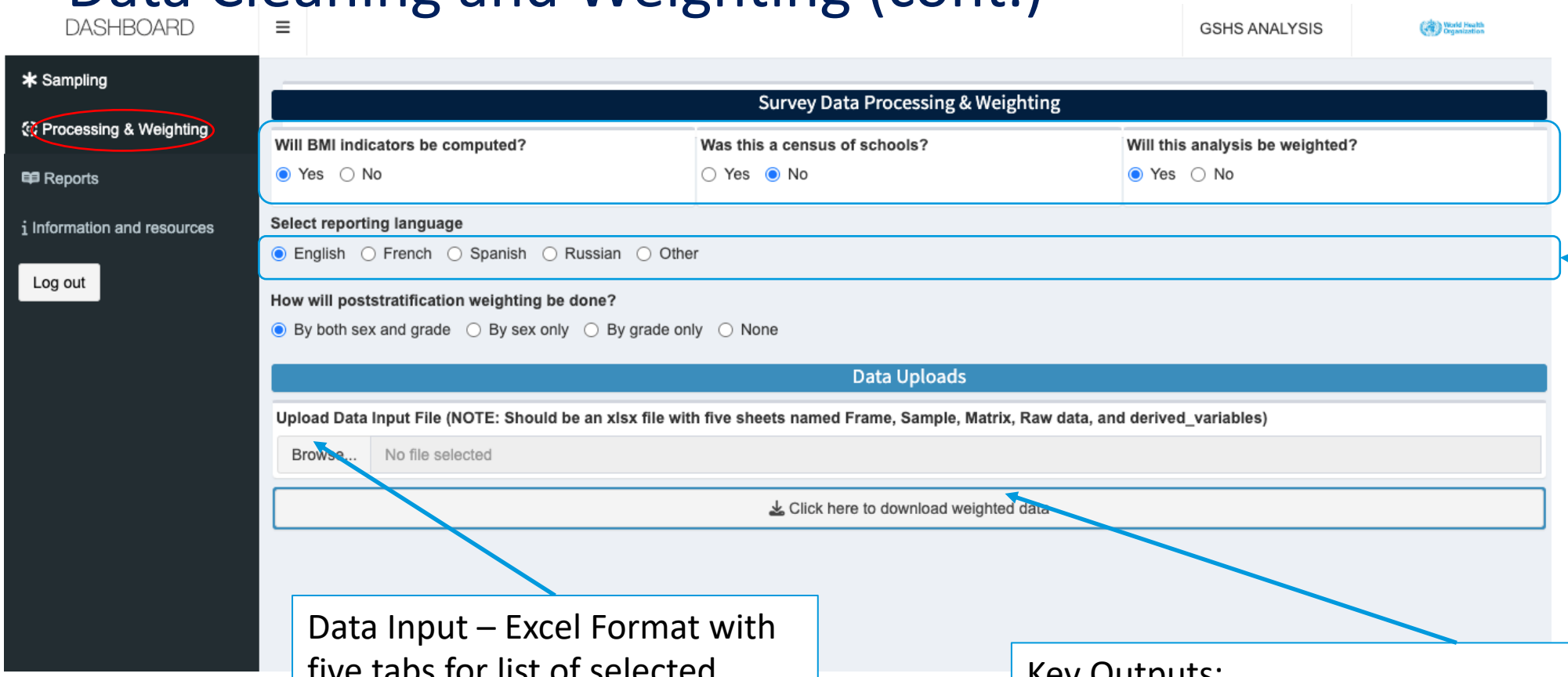
Variables needed to derive secondary variables

Logical condition for the numerator

Logical condition for the denominator

World Health Organization

Data Cleaning and Weighting (cont.)



Controls for reporting BMI indicators, census, and weighted analysis

Control for languages – English, French, Spanish, and Russian

Data Input – Excel Format with five tabs for list of selected schools, frame, weighted data, indicator matrix, and derived variables

Key Outputs:
An Excel file with the following:

- Two cleaned weighted and standardized datasets
- Updated Matrix
- Frame
- Sample
- Derived variables tab

World Health Organization

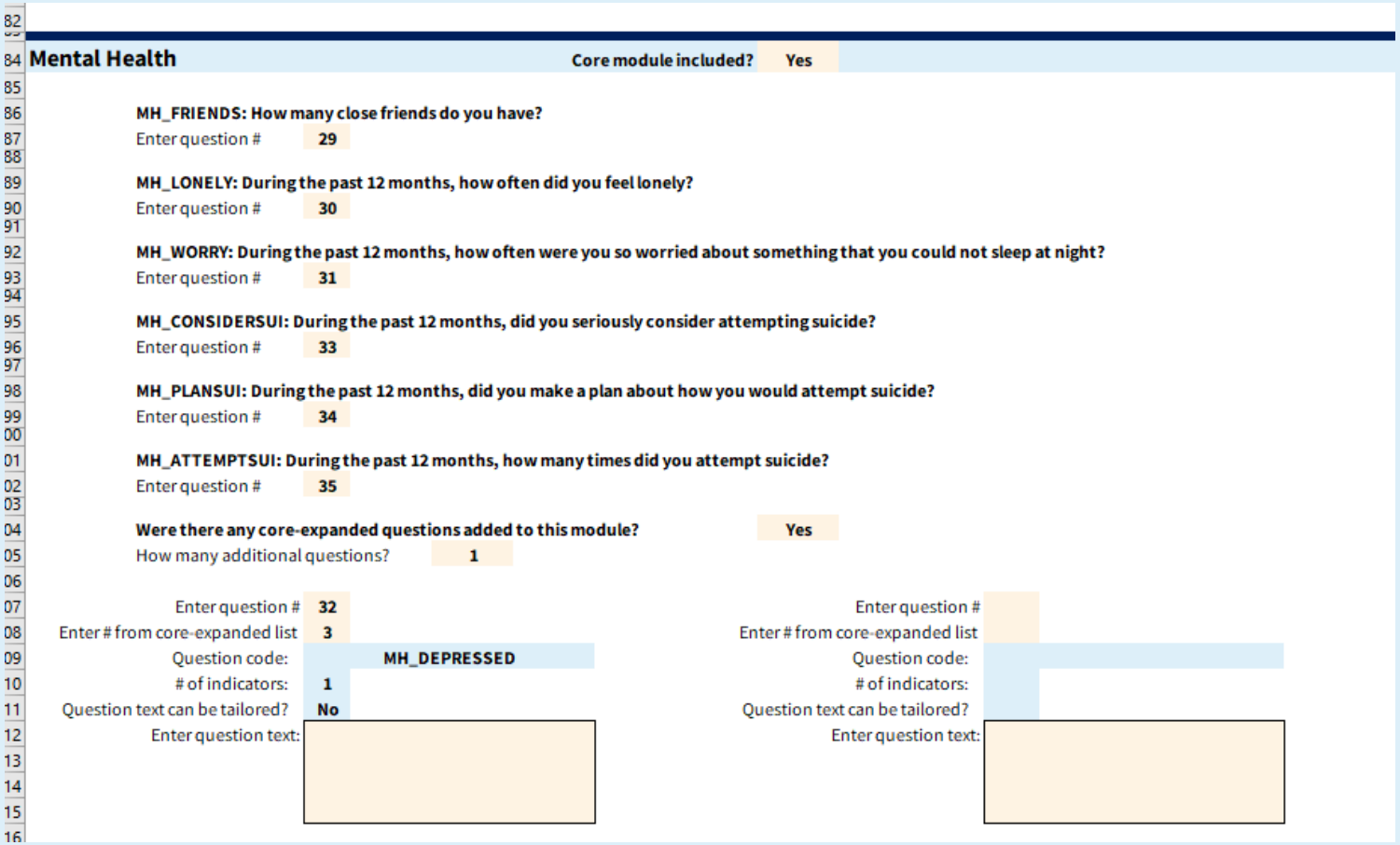
Tools to help

- There is a standard **log-in form** for GSHS which, once completed, will create the **sample.xlsx** file.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	SITE NAME 2023 GSHS															
2								Participated	Classes		Class ID			Class ID		
3								1=yes/0=no	Total #	# Classes	Class	Total #	# Students	Class	Total #	# Students
4	Number	School ID	School	Enrollment	Category	School Weight	School Interval	Ineligble=Blank	Classes	Selected	Selected	Enrolled	Participated	Selected	Enrolled	Participated
5	1	1	A	1142	national	1	2.929408	1	22	8	1	36	19	4	31	10
6	2	2	B	1020	national	1	2.929408	1	19	6	2	39	29	5	35	30
7	3	3	C	787	national	1	2.929408	1	20	7	2	39	29	5	31	19
8	4	4	D	634	national	1	2.929408	1	14	5	1	36	28	4	33	14
9	5	5	E	509	national	1	2.929408	1	13	4	2	88	65	5	40	40
10	6	6	F	508	national	1	2.929408	1	10	3	2	28	28	5	34	34
11	7	7	G	401	national	1	2.929408	1	9	3	1	32	27	4	32	26
12	8	8	H	399	national	1	2.929408	1	12	4	1	35	21	4	33	28
13	9	9	I	311	national	1.215288282	2.410463463	1	12	5	2	32	24	4	37	28
14	10	10	J	290	national	1.303291916	2.247699049	1	6	3	1	32	29	3	28	22
15	11	11	K	279	national	1.354676185	2.162441499	1	10	4	2	30	29	4	32	27
16	12	12	L	276	national	1.369400926	2.13918944	1	10	5	1	7	7	3	27	25
17	13	13	M	264	national	1.431646423	2.046181203	1	7	3	2	55	36	4	50	36
18	14	14	N	254	national	1.488010455	1.96867434	1	6	3	2	20	16	4	22	20
19	15	15	O	239	national	1.581400233	1.852414044	1	11	5	2	28	26	4	43	35
20	16	16	P	215	national	1.757928631	1.666397571	1	10	6	1	34	32	2	30	30
21	17	17	Q	208	national	1.81708969	1.612142766	1	4	2	1	38	34	2	37	24
22	18	18	R	182	national	2.076673932	1.41062492	1	4	2	1	37	31	2	37	33

Tools to help

- There is also a **matrix generation tool** which allows you to enter information about your questionnaire and then use a macro to generate your **mapping_matrix.xlsx** file.



Mental Health — Core module included? Yes

MH_FRIENDS: How many close friends do you have?
Enter question # 29

MH_LONELY: During the past 12 months, how often did you feel lonely?
Enter question # 30

MH_WORRY: During the past 12 months, how often were you so worried about something that you could not sleep at night?
Enter question # 31

MH_CONSIDERSUI: During the past 12 months, did you seriously consider attempting suicide?
Enter question # 33

MH_PLANSUI: During the past 12 months, did you make a plan about how you would attempt suicide?
Enter question # 34

MH_ATTEMPTSUI: During the past 12 months, how many times did you attempt suicide?
Enter question # 35

Were there any core-expanded questions added to this module? Yes
How many additional questions? 1

Enter question #	32	
Enter # from core-expanded list	3	
Question code:	MH_DEPRESSED	
# of indicators:	1	
Question text can be tailored?	No	
Enter question text:		

Enter question #	
Enter # from core-expanded list	
Question code:	
# of indicators:	
Question text can be tailored?	
Enter question text:	

World Health Organization