EDUCATION LEVELS - General Target Variable Report (GVR)

1. General Information

T_EDU is a target variable that categorizes respondent's level of education into three-digit codes.[1] The first digit distinguishes between no formal education, primary, secondary or tertiary level of education. The second digit indicates the sub-level of education within primary, secondary and tertiary, respectively. The third digit distinguishes between incomplete and complete education, whenever the source question explicitly specifies that a given education level is incomplete (see Table 1.1). The order of education levels in the target variable corresponds to the International Standard Classification of Education levels (ISCED 2011), but the specific SDR2 codes differ .

The harmonized variable T_EDU is accompanied by the following harmonization control variables (see Section 3.3):

- C_EDU_VOCATIONAL: educational level has vocational or technical component,
- C_EDU_SOURCE: target variable is based on a country-specific or standardized classification of education,
- C_EDU_CATEGORY_NUM: provides the total number of education categories in a source schema,
- C_EDU_CATEGORY_OPEN: flags the situation when a source category uses modifiers "less" or "higher" to describe a given education level.

The target variable report for T_EDU is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR) T_EDU_DVR_SDR2.xlsx. DVR Excel files in SDR2 systemize all information about source variables that were used for harmonization into a given target variable of the SDR2 database;
- Three Crosswalk Tables (CWT): T_EDU_CWT_SDR2.xlsx, T_EDU_VOCATIONAL_CWT_SDR2.xlsx and T_EDU_CWT_CATEGORY_OPEN_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

  T_EDU_CWT_SDR2.xlsx file contains details about mapping of source values to target to create T_EDU, C_EDU_SOURCE and QP_EDU control variables.

  Additional two CWT files - T_EDU_VOCATIONAL_CWT_SDR2.xlsx and T_EDU_CWT_CATEGORY_OPEN_SDR2.xlsx – capture codes for two control variables: T_EDU_VOCATIONAL and T_EDU_CATEGORY_OPEN, accordingly.

[1] The SDR2 project also harmonizes a target variable, T_EDU_YEARS, which captures the number of years of education completed by a respondent.

Table 1.1. EDUCATION LEVEL: Description of the target, source, and control variables

	Variable description	**Variable name**	**Variable values***
Target variable	Education category capturing levels, sublevels and complete/incomplete status of respondents' highest formal education	T_EDU	0 No formal education 101 Primary incomplete 102 Primary complete/unspecified 111 Lower secondary incomplete 112 Lower secondary complete/unspecified 201 Secondary incomplete 202 Secondary complete/unspecified 211 Upper secondary incomplete 212 Upper secondary complete/unspecified 221 Post-secondary non-tertiary incomplete 222 Post-secondary non-tertiary complete/unspecified 301 Short-cycle tertiary incomplete 302 Short-cycle tertiary complete/unspecified 311 Higher/tertiary education incomplete 312 Higher/tertiary education complete/unspecified 321 BA or equivalent incomplete 322 BA or equivalent complete/unspecified 331 MA or equivalent incomplete 332 MA or equivalent complete/unspecified 341 PhD or equivalent incomplete 342 PhD complete/unspecified 400 Still at school as a response option
Source variables			See: T_EDU_DVR_SDR2.xlsx T_EDU_CWT_SDR2.xlsx T_EDU_CWT_VOCATIONAL_SDR2.xlsx T_EDU_CWT_CATEGORY_OPEN_SDR2.xlsx

Control variables	Education level with vocational or technical component (inferred from source documentation)	C_EDU_VOCATIONAL	0 No sufficient information to determine if category is vocational 1 Category is vocational 2 Category is not vocational
	Classification type of source education schema	C_EDU_SOURCE	1 Target variable based on country-specific classification of education 2 Target variable based on standardized classification of education
	Total number of education categories in source schema	C_EDU_CATEGORY_NUM	2 Two categories in a source variable … 42 Forty-two categories in a source variable
	Source schema contains open category	C_EDU_CATEGORY_OPEN	0 Exact level 1 Category is open to the bottom (e,g, secondary education and less) 2 Category is open to the top (e.g. bachelor degree and higher)
Quality control (stored in PLUG_SURVEY file)	Quality control: Processing error types (realized values)	QP_EDU	0 No processing errors 2 Misleading variable values 3 Contradictory variable values 5 Lack of variable value labels 10 Contradictory variable values & lack of variable value labels

* Missing values are assigned according to the SDR2 missing codes schema, provided in the Appendix.

2. Survey Projects

Source variables that we used for T_EDU appear in 23 international survey projects: ABS, AFB, AMB, ARB, ASES, CB, CDCEE, CNEP, EB, EQLS, ESS, EVS, ISJP, ISSP, LB, LITS, NBB, NEB, PA1, PA2, PPE7N, VPCPCE, WVS, 127 waves and 2289 national surveys. The data cover 148 countries and years from 1966 to 2017.

3. General Rules and Procedures

3.1. Source data description

International survey projects often ask about respondents' level of education using similarly worded questions. However, the possible response options and the categorization of education can vary. We outline the variations in the following issues:

- International vs country-specific classification: Some survey waves use national (country-specific) classification of education, while others use international categorization (like ISCED), or their own categorization applied to all countries within a survey wave. We created a control variable C_EDU_SOURCE to capture this information.

- Additional information: Each source variable that we use includes information on the highest completed education level. However, we sometimes encounter additional information, such as whether respondents started but not finished a particular level of education (complete vs incomplete education level), or if a response option includes a vocational component (e.g. technical college). We capture this information through the third-digit code of our target variable and construct a control variable C_EDU_VOCATIONAL, accordingly.

- Merged categories: Some surveys combine multiple education outcomes into one response option, such as *MA and higher*, which includes both MA and PhD degrees; or *primary or incomplete secondary*, which includes both primary education and some years of secondary education. We constructed two control variables to capture these situations: C_EDU_CATEGORY_NUM, which indicates the number of response categories within one source variable and C_EDU_CATEGORY_OPEN, which flags the first or the last source category that merges two or more levels of education into one response option.

3.2. Rules of transformation of source variables into target variable

In the absence of any officially recognized education level, we coded T_EDU as 0, no formal education, which constitutes the lowest level of education in the SDR2 data.

The third digit of the T_EDU code indicates whether an education level is incomplete or complete. Specifically:

1. If the source documentation explicitly states that a certain level of education is incomplete, we code T_EDU = "1" in the third digit.
2. If the source documentation explicitly states that the education level is complete, we code "2" in the third digit. This is often the case within source variables that distinguish between incomplete and completed education levels.
3. If the source question asks about the highest completed level of education and the source variable does not provide the answer category "incomplete," we assume that the education level is completed and code "2" in the third digit.

For primary education - reflected in the source labels “primary”, “elementary” and “basic” education , we use the codes ‘101’ and ‘102’, where the first two digits (10) indicate that the education level is primary, while the third refers to incomplete/completed. If a source variable does not include the response category “incomplete primary,” but only completed levels of education, and “no formal schooling” is followed by “primary level,” we code “no formal schooling” as 101, assuming that it also captures “incomplete” primary.

For lower secondary education, we use the codes ‘111’ and ‘112’, where the first two digits (11) indicate that the level is lower secondary. This level includes categories such as lower/junior secondary, grammar school, vocational, occupational school and others.

Secondary education levels beyond lower secondary are captured by codes starting at ‘201’. This includes general secondary education (coded as ‘201’ and ‘202’), which encompasses secondary, upper middle secondary and some vocational categories of education.

Upper secondary education is represented by codes ‘211’ and ‘212’, which cover higher / upper secondary, vocational education.

Post-secondary non-tertiary education is represented by codes ‘221’ and ‘222’, which include post-secondary education, technical and professional schools, semi-higher education, and other categories.

Tertiary education is categorized into three sub-levels, with short-cycle tertiary education coded as ‘301’ and ‘302’ (e.g. junior college, some college, community college, associate degree in college) and higher tertiary education without distinction coded as ‘311’ and ‘312’ (when it is possible to establish that a higher level of education completed is tertiary, but it is not possible to establish what level of tertiary education was completed exactly). Bachelor degree level (BA) or equivalent is coded as ‘321’ and ‘322’, Master degree level (MA) or equivalent is coded as ‘331’ and ’332’, while PhD level or equivalent is coded as ‘341’ and ‘342’. Note that none of the source variables used to create the target variable contained information that would have allowed us to code the 341 category for “PhD or equivalent, incomplete”.

When mapping source values into the target, we take into account the entire set of response categories for a variable, rather than relying solely on the textual labels. This is because the labels often reflect local educational contexts, which can vary widely across countries and regions. For example, while college education might be the highest level of education in some countries, it could be considered a third-cycle tertiary degree elsewhere. Furthermore, response categories can differ in their level of detail. As a result, the same textual label (e.g. “college education”) might receive different codes in the SDR2 target depending on the context and accounting for the entire set of response categories.

If a respondent is still at school/college/university/training, we code it as T_EDU = 400.

In education systems that recognize leaving certificates (such as the UK, France), the SDR2 level of education corresponds to a specific certificate that represents the highest level of education attained. If a respondent completed education without receiving a certificate, the level is coded as incomplete.

In cases where special education is pursued beyond primary schooling, such as nursing schools or agricultural schools, the corresponding education level is coded with a control variable C_EDU_VOCATIONAL = 1, indicating vocational or technical training. Further details about this and other control variables can be found in section 3.3 of this report.

Missing values and different situations that warrant to be treated as missing data are coded according to the SDR2 missing codes schema, provided in Table A.1 in the Appendix.

3.3. Methodological variables that accompany T_EDU

We constructed four harmonization control variables for T_EDU:

C_EDU_VOCATIONAL is a nominal variable that indicates vocational or technical training for any type of special education above the primary level (102 code) and up to (and including) post-secondary non-tertiary level (222 code). This variable takes the value '0' if it is not possible to establish from the available source documentation if the education category is vocational or technical.

If the source documentation indicates a vocational or technical component for an educational category, such as a direct in-text label "vocational" or specific labels such as nursing schools or agricultural schools, C_EDU_VOCATIONAL takes the value "1." This code also flags responses that merge categories of vocational/technical and general education, such as "general plus vocational training" or "vocational plus maturity" (see for example ISSP_1988 Netherlands v70).

C_EDU_VOCATIONAL takes the value 2 when a source education category is not vocational or technical. We use this code only when a source variable differentiates between educational levels categorized as vocational/technical and those categorized as general/not vocational. For example, if a source variable has response categories such as "… lower secondary vocational, lower secondary not vocational, upper secondary, …" , then 'lower secondary vocational' would be coded as '1', lower secondary not vocational' as '2' and 'upper secondary' as '0' . This is because the source variable does not distinguish between vocational and non-vocational education at the upper secondary level.

C_EDU_SOURCE indicates the type of education schema used in the source variable. When the source variable relies on a country-specific classification of education, C_EDU_SOURCE takes the value 1. We code this control as '2' when the harmonized source variable uses a standardized classification of education, with education categories common for all countries of a source

project's wave. Typically, survey projects that use standardized classification have one source variable that captures education levels across all countries of that project, although sometimes there is a set of national variables that uses that standardized classification. Even if there is a set of different variables for countries within a project wave, but they all share the same (identical) schema of coding education, we code C_EDU_SOURCE as '2'. For instance, this is the case for the ARB_3 and VPCPCE project waves.

C_EDU_CATEGORY_NUM indicates the total number of education categories in a given source schema. The total number of education categories of a given source variable in SDR2 ranges from 2 (CDCEE_1 Lithuania) to 42 (PA1 Netherlands) response options per one source variable (based on realized values). Note that this total number includes not only education levels, sublevels and complete/incomplete categories, but also categories such as "still at school" and "no education" if they appear as response options in the source variable.

C_EDU_CATEGORY_OPEN flags the situation when a source category uses the modifiers "less" or "higher" in the description of a given education level, thus opening that level to include respondents with lower or higher education. We use the code '1' when the source category label specifies a level and "less," (e.g., "secondary education and less"). A code of '2' means that a category includes a level and "'higher" (e.g., "bachelor degree and higher"). We use the code '0' for categories that correspond to an exact level.

This control variable identifies categories in source schema that combine two or more levels of education. For example, the source category 'secondary education and less' comprises the secondary and primary levels of education, as well as no formal education. When coding C_EDU_CATEGORY_OPEN, we take into account the context of all response categories. For example, if the last response category is "Master degree" and no other "higher" education level responses are available (as in ASES V0233), we assume that people with PhD diplomas are also captured by this category. Therefore, we code it the same way as when the label is "tertiary and higher".

3.4. Quality Control Variable for processing error types

QP_EDU, stored in the SDR2 PLUG_SURVEY file, measures processing error types that we identified when comparing information about the source variables available in the survey documentation (codebook, questionnaire) with information about the same source variables in data records in the source data files (data dictionaries). QP_EDU accounts for inconsistencies between different metadata elements, such as data records on the one hand (i.e. variable values, variable labels, value labels – in the source data files), and information in the codebook or questionnaires, on the other hand.

QP_EDU is a nominal variable that takes values from 0 (no processing errors) to 13. Values larger than 0 identify what type, or combination of types, of processing errors occur in the source variables associated with T_EDU.

Generally, in the SDR2 project we check for five types of processing errors that we define as follows:

- Illegitimate Variable Values flags values of the source variable that are outside of the acceptable range proposed by SDR. For example, a number of formal education years exceeding 25 or an 18-year-old respondent reporting PhD as their education level would be considered illegitimate source values. We did not identify any such errors in the source variables on education levels.
- Misleading Variable Values shows that variable values, as coded in the questionnaires/codebooks (e.g., education levels), are not congruent with the data records in the source data files.
- Contradictory Variable Values captures inconsistencies in how the same variable is labeled or the variable values in a codebook vs. a questionnaire or a data dictionary. E.g. when the question asks about the number of years of education, while the answer options suggest that the question deals with education levels.
- Variable Values Discrepancy concerns situations when several or all values in the same variable are inconsistently labeled in codebook, vs. questionnaire or a data dictionary (i.e. the scale differ or contradictory value labels for more than one variable value appear).
- Lack of Variable Value Labels concerns undefined "nulls" and codes (variable values) in the source data file that are not explained in any source of documentation defining variables and their values (i.e. "wild codes" for a given variable, where values in data records exceed the values given in documentation).

Correspondingly, the full list of SDR2 codes for Processing Error Types is::

0 = No processing errors
1 = Illegitimate variable values
2 = Misleading variable values
3 = Contradictory variable values
4 = Variable values discrepancies
5 = Lack of variable value labels
6 = Illegitimate & misleading variable values
7 = Illegitimate & contradictory variable values
8 = Illegitimate & discrepancy variable values
9 = Illegitimate & lack of variable value labels
10 = Contradictory variable values & lack of variable value labels
11 = Discrepant & lack of variable value labels
12 = Illegitimate & contradictory & lack of variable value labels

13 = Illegitimate & discrepant & lack of variable value labels

Values 6 to 13 indicate that a combination of different processing error types occurs within one source variable.

The source variables that we used to create a target T_EDU had misleading variable values (QP_EDU = 2), contradictory variable values (QP_EDU = 3) and lack of variable value labels (QP_EDU = 5) or a combination of these processing error types (QP_EDU = 10). Other types of processing errors did not occur in source data used for this target variable, hence the realized values in the QP_EDU variable are 0, 2, 3, 5 and 10.

4. Special Cases

- EQLS_1_3 survey project, variable Y11_ISCED includes a source category "completed education abroad." We code it as T_EDU -8 (.h) "unfit," which is the SDR2 missing code for source values that do not match our target. See details on SDR2 missing code schema in Table A.1 of the Appendix.
- LB_2005 *s11* education source variable captures education in years for some categories and education levels (complete/incomplete) for other categories within the same variable. We map education in years into education levels for this variable. Details are available in T_EDU_CWT_SDR2.xlsx file.
- To code education level corresponding to the answer option "vocational," we often had to rely on the context of other response options in which this specific category appears. For example, in CDCEE_1_2 v597 categories for the level of education are "… elementary, primary, vocational, secondary, college, higher ….". Taking into account that the response option "vocational" appears after "primary" and before "secondary", we assume that it refers to secondary vocational training and assign it the value "202." For details see T_EDU_CWT_SDDR2.xlsx and T_EDU_CWT_VOCATIONAL_SDR2.xlsx files.

- In AMB_2006, we used *ed* variable to code respondent's education level for Canada and the United States. We were not able to map the education level for other countries from AMB_2006 based on the source documentation.

  In AMB_2008, we used *ednivel* source variable for all countries but Canada. To code Canada from AMB_2008 we used an additional variable *ed.*

  In AMB_2016 we used *ed_can* variable to code education level for Canada and *ed_usa* variable for the United States.

- For ISSP and ESS waves 1- 4, we used the aggregated level variable to code the levels of education, although more detailed, national level codes are available in the source data.
- In the cases of
  - EB_0_2 v120 for France, Denmark and Great Britain, where the source label says "High school, University"
  - NEB_7 G5SLE variable, answer category "High, college or more" and
  - PA2 v1199 variable, answer category "University, teachers' training college",
  source labels suggest that both university education level and, possibly, short-cycled tertiary levels are merged within one response category. However, based on the context in which these response categories appear, we decided to code them as 312 "higher education complete" with the control category open = 2 "and higher" as this is the last response category from the list of source options and, in addition, may include respondents with e.g. PhD level.

5. Additional comments

Our main additional source is the document on the International Standard Classification of Education ISCED 2011 issued by UNESCO: http://www.uis.unesco.org/Education/Documents/isced-2011-en.pdf

Appendix: Codes for missing values in SDR2

In the SDR database v.2 we identify different situations that warrant to be treated as missing data. Table A.1 lists all SDR2 missing value codes:

Table A.1. Codes for missing values in SDR2

SDR tag [a]	SPSS (STATA) codes	Label
		Standardized source codes for missing values
DK	-1 (.a)	Don't know
NA	-2 (.b)	No answer
REF	-3 (.c)	Refusal
DU	-4 (.d)	Don't understand the question
DNR	-5 (.e)	Any combination of DK, NA, REF, DU
INAP	-6 (.f)	Inapplicable
NEC	-7 (.g)	Not elsewhere classified
		SDR created codes for missing values
UNFIT	-8 (.h)	Source value does not fit to target
ERR	-9 (.i)	Errors in source data and undocumented source values
COMBI	-10 (.j)	Different missing codes on multiple sources taken for a target
CINAP	-11 (.k)	For control variables only: inapplicable
INSUF	-12 (.l)	For survey: Insufficiently defined response categories
QNA	-13 (.m)	For survey: Question not available

[a] Abbreviations for the labels corresponding to the SDR2 codes for missing values. These tags are used in the Cross-walk Table (CWT) files (Excel) that accompany documentation of SDR2 target variables.

In exceptional situations when codes for missing data listed in Table A.1 cannot be used, we apply a system missing <null> value.