## YEARS OF EDUCATION - General Target Variable Report (GVR)

### 1. General Information

The target variable T_EDU_YEARS measures respondents' completed years of education[1]. The measure takes the value 0 when respondents report that they do not have formal education. It takes values 1 to 22 indicating the number of education years completed, where 22 means 22 years or more (see Table 1.1). The assumption behind cutting the upper boundary at 22 is that secondary education on average lasts for 12 years, higher education for another six years, and four more years for graduate and postgraduate formal education (cf., OECD, European Union, UNESCO Institute for Statistics 2015; Whipple 2016).

The target variable T_EDU_YEARS is accompanied by three harmonization control variables specifying: (a) if the variable was derived (C_EDU_YEARS_COMPUTED); (b) if the value range is open (C_EDU_YEARS_OPEN); c) if the values exceed 22 years of education (C_EDU_YEARS_ABOVE_22); as well as by harmonization quality control QP_EDU_YEARS measuring the occurrence of processing errors on source variables used to construct T_EDU_YEARS (see Table 1.1 and Sections 3.3 and 3.4).

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

- The Detailed Variable Report (DVR) T_EDU_YEARS_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;
- The Crosswalk Table (CWT): T_EDU_YEARS_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

**Table 1.1. YEARS OF EDUCATION**: Description of the target, source, and control variables

| | **Variable description** | **Variable name** | **Variable values[a]** |
|---|---|---|---|
| **Target variable** | Completed years of education | T_EDU_YEARS | 0 = No formal education<br>1 = 1 year<br>22 = 22 years or more |
| **Source variables** | | | See: T_EDU_YEARS_DVR_SDR2.xlsx and |

[1] SDR2 project also harmonizes a target variable T_EDU that captures respondent's education into three-digit target categories.

| | | | |
|---|---|---|---|
| | | | T_EDU_YEARS_CWT_SDR 2.xlsx |
| **Control variables** | Years of education computed by survey provider or SDR team from other variables | C_EDU_YEARS_COMPUTED | 0 = No transformation made - source values mapped to target (raw values)<br><br>1 = Computed by SDR team from age when R finished education<br><br>2 = Computed by survey providers from age when R finished education<br><br>3 = Computed by survey provider from R's level of education<br><br>4 = Computed by survey provider using other information<br><br>5 = Impossible to identify whether the value was derived (no documentation)<br><br>6 = Computed by SDR team from "still at school" category |
| | Ends of source value range for education years' are open | C_EDU_YEARS_OPEN | 1 = Lowest end of range is open<br><br>2 = Highest end of range is open<br><br>3 = Both ends of range are open<br><br>0 = Otherwise |

| | | | |
|---|---|---|---|
| | Source value of education years range is cut at 22 years | C_EDU_YEARS_ABOVE_22 | 1 = Source values exceed 22 (23 and more)<br><br>0 = Otherwise |
| **Quality control variables for processing error types** | Processing error types (realized values) | QP_EDU_YEARS | 0 = No processing errors<br>1 = Illegitimate variable values<br>3 = Contradictory variable values<br>4 = Discrepant variable values<br>5 = Lack of variable value labels<br>7 = Illegitimate & contradictory variable values<br>8 = Illegitimate & discrepant variable values<br>9 = Illegitimate & lack of variable value labels<br>10 = Contradictory variable values & lack of variable value labels<br>11 = Discrepant & lack of variable value labels<br>12 = Illegitimate & contradictory & lack of variable value labels<br>13 = Illegitimate & discrepant & lack of variable value labels |

[a] 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_YEARS appear in 20 international survey projects: ABS, AMB, ASES, CB, CDCEE, CNEP, EB, EQLS, ESS, EVS, ISJP, ISSP, LB, LITS, NBB, NEB, PA1, PPE7N, VPCPCE, WVS, 152 waves and 2904 national surveys. The data cover 131 countries and years from 1971 to 2017.

## 3. General Rules and Procedures

### 3.1. Source data description

To construct the target variable T_EDU_YEARS, we use source items about respondents': (a) exact number of education years completed; (b) age when finished full-time

education; (c) year when finished school; (d) years of education derived by survey providers from various source variables. We rely on the English language and Spanish language questionnaires and codebooks describing the source survey data.

Typical questions on respondent's years of education are: "*How many years of formal education have you received?*" (ABS), "*How many years of schooling have you completed?*"(AMB/2010-2016), "*About how many years of education have you completed, whether full-time or part-time?*" (ESS). Typical questions on the age of completing education are: "*How old were you when you finished your full-time education?*" (CDCEE), "*At what age did you finish your education (full-time education)?*" (LB). In two cases, respondents were asked to indicate the year of their education completion: *"When did you obtain this degree?"* (LITS/1), *"When did you obtain this qualification?"* (LITS/2).

### 3.2. Rules of transformation of the source variable into target variables

For ex-post harmonization, we select four types of source variables that refer to respondents' years of education.

1. The first type of our source variables provide information about years of education directly. It is either based on the source question, which asks about years of education or on the source variable, which contains the number of years as computed by the survey provider. For this type of source variables we preserve the maximum of source data without any specific modifications (for a few special cases: see Section 4).

2. The second type of source variables contain information about respondents' age of completing full-time education. In this case, we subtract 6 years from respondents' age to achieve the value of completed education years. The assumption behind this decision is that, generally, primary education begins at the age of 6 years. If subtracting 6 years results in negative values, such values are treated as invalid, i.e., missing data and coded as ERR (errors in source data and undocumented source values, see Table A.1 in the Appendix). If the result of subtraction is 0, we code it as 0, i.e., "no formal education".

When source variables indicate the year of the education completion, first the age of respondents in a given year is established, then the number of education years is calculated by subtracting 6 years from respondents' age.

If source items include "still at school" response option, the number of education years is calculated by subtracting 7 years from responts' age assuming that such respondents are continuing their education one level above the completed one.

In case of incomplete years in source data e.g., "10.5 years" (as happens in ABS and ESS), the number is rounded to the lower complete year.

Whenever the range of years appears in source variables labels, we follow source values given by the survey provider.

We set the upper boundary of education years to 22 and create a category “22 and more”. 22 years of education corresponds to the PhD level, the highest degree of education. This number follows from (approximate) cumulative counting of years of schooling for (consecutive) levels of education (cf., OECD, European Union, UNESCO Institute for Statistics 2015; Whipple 2016). We mark all cases when more than 22 years of education are reported in source data with the help of a control variable C_EDU_YEARS_ABOVE_22.

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_YEARS.

The target variable T_EDU_YEARS is accompanied by three harmonization control and one quality control variables. All methodological variables are coded at the country level.

C_EDU_YEARS_COMPUTED specifies whether the source or the target variable was computed by the survey provider or by the SDR team. It takes the value 0 if the variable is not computed. It takes the value 1 if target values are derived by the SDR team from age when respondents completed education. It takes the value 2 if source values are computed by the survey provider using respondents’ age when finished school. It takes the value 3 if source values are calculated by the survey provider using respondents’ levels of education. It takes the value 4 if source values are obtained by the survey provider from a combination of source items (e.g., ISSP/2004, DE-W: *Constructed from 'general school leaving certificate', 'vocational training' and 'details of non-employment'*). It takes the value 5 if the information is not documented and it is impossible to identify whether source values were derived. It takes the value 6 if target values are computed by the SDR team from the source category “still at school”. See more details in Section 4.

C_EDU_YEARS_OPEN indicates whether the response option is “open”, i.e., a given source value signifies a number of years “and less” or “and more” (e.g., EB/5: 1 = *up to 14 years*, 22 = *22 years or older*)”. If the lower edge of the range is open, we assign the value 1. If the upper edge is open, we assign the value 2. If both edges are open, we assign the value 3. Otherwise, C_EDU_YEARS_OPEN takes the value 0.

C_EDU_YEARS_ABOVE_22 marks source values that exceed 22 years of education (23 years and more). In this case, the range is cut at 22 years by the SDR team and C_EDU_YEARS_ABOVE_22 takes the value 1. Otherwise, it takes the value 0.

### 3.4. Quality Control Variable for processing error types

QP_EDU_YEARS, 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). Put differently, QP_EDU_YEARS 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_YEARS is a nominal variable that can take 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_YEARS.

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

- **Illegitimate Variable Values** captures values of the source variable that are outside of the range that SDR proposes as acceptable, e.g., respondents' number of education years exceeding 25, and age when full‑time education was completed below 6.

- **Misleading Variable Values** shows that variable values, as coded in the questionnaires/codebooks, are not congruent with the data records in the source data files. T_EDU_YEARS, however, does not contain such errors.

- **Contradictory Variable Values** captures inconsistencies in how the same variable value is labeled in a codebook vs. a questionnaire or a data dictionary, such as when "don't know" value is recorded as 88 in the codebook vs. 888888 in the data dictionary.

- **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., there are reversed scales or contradictory value labels for more than one variable value), e.g., age of responds when finished education is recorded as year ranges in the questionnaire vs. as years in the data dictionary.

- **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, such as value 7 in data records while the documentation describes the variable as a 1 to 5 scale).

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.

## 4. Special Cases

- ISSP/2010 `educyrs`: Scale for New Zealand and Austria differ from other countries. In particular the control variables C_EDU_YEARS_OPEN and C_EDU_YEARS_COMPUTED were set to 3.
- In the case there is no information in source documentation as to how the source values of years of education were obtained, we look at the distribution and/or at other waves for the same country. If the distribution appears to be “chopped”, we assign the value 3 to C_EDU_YEARS_COMPUTED, since it is clear that these values were derived from education levels (e.g., ISSP/1993,1994, Ireland). When the distribution is continuous, we assign the value 5 to C_EDU_YEARS_COMPUTED (e.g., ISSP/1992,1996 Hungary). If all the waves except for those undocumented have the same repeated source question and in this sense are standardized, we assign the same value for all the waves for this country (e.g., ISSP/1997,1999,2001,2003, Great Britain).
- In EQLS/1-2 `Y07_CVq48`, the option "never completed" coded as 97 or 997 was recoded to 0 (no formal education).

## References

OECD, European Union, UNESCO Institute for Statistics. (2015). ISCED 2011 O*perational Manual: Guidelines for Classifying National Education Programmes and Related Qualifications.* OECD Publishing. http://dx.doi.org/10.1787/9789264228368-en.

Whipple, S. [Ed.] (2016). *A Guide to Educational Systems Around the World. Revised Edition.* Washington, DC: NAFSA Association of International Educators. PDF on CD, Item CD2016.

**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 | -20 (.t) | For survey: Question not available |

[a] Abbreviations for the labels corresponding to the SDR2 codes for missing values. These tags are used in the Crosswalk 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.