SUBJECTIVE HEALTH ASSESSMENT – General Target Variable Report (GVR)

1. General Information

We provide two harmonized measures of respondents' self-assessed general health: T_HEALTH_5 and T_HEALTH_DISTRIB. Section 3.2 outlines the harmonization rules for each target variable.

T_HEALTH_5 measures respondents' subjective health assessment on a 5-point scale. Target values range from 0 (worst) to 4 (best) health.

T_HEALTH_DISTRIB measures the relative position of a respondent in the distribution of self-assessed health in a given national survey. The scores of this target variable are percentiles within the national sample that indicate what share of respondents reports the same or poorer health than the individual.

Both target variables are accompanied by five harmonization controls that capture specific features of the "health assessment" source variables (see Table 1.1, and Section 3.3):

1. C_HEALTH_PHYS is dichotomous, showing that the source question asks specifically about respondents' physical health (value 1), or their health in general (value 0).
2. C_ HEALTH_TIME is a nominal variable that indicates whether the respondent's state of health is assessed without reference to any specific time frame (value 0), with reference to the time around the interview (value 1), or with reference to the past year (value 2)
3. C_HEALTH_LENGTH is a nominal variable that identifies the length (i.e. number of answer options) of a given source scale used to construct T_HEALTH_5 and T_HEALTH_DISTRIB.
4. C_HEALTH_FAIR is a nominal variable. It takes the value 0 if "fair" does not appear among a scale's answer options in the source master questionnaire (in English). It takes the value 1 if "fair" constitutes the midpoint of odd-numbered source scales (i.e. "fair" corresponds to average health). When "fair" appears as an answer option that should capture below-average health (e.g., in a 5-point scale with response options: excellent (1); very good (2); good (3); fair (4); worst (5)), we assign C_HEALTH_FAIR the value 2.
5. C_HEALTH_ASCEND is dichotomous. It takes the value 1 for source scales whose values are in ascending order (i.e. response options are ordered from worst to best health); it takes the value 0 when source scale values are in descending order (i.e. response options are ordered from best to worst health).

The target variable report for subjective health assessment is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR): T_HEALTH_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_HEALTH_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

Table 1.1. SUBJECTIVE HEALTH ASSESSMENT: Description of the target, source, and control variables

	Variable description	Variable name	Variable values[a]
Target variable	Subjective health assessment (5-point scale)	T_HEALTH_5	0 = worst health … 4 = best health
	Subjective health assessment (distribution- preserving scale)	T_HEALTH_DISTRIB	0 = lowest percentile point in the distribution 100 = highest percentile point in the distribution
Source variables	Source values of Respondent's subjective health assessment		See: T_HEALTH_DVR_SDR2.xlsx T_HEALTH_CWT_SDR2.xlsx
Control variables	Type of health the source question(s) ask about	C_HEALTH_PHYS	0 = source variable asks about health in general 1= source questions ask about physical health
	Timeframe for health assessment referenced in the source variable	C_HEALTH_TIME	0 = source question does not reference any time frame 1 = source question explicitly asks about respondent's health around the time of interview (e.g. at the moment/now/these days) 2 = source question asks about the respondent's health in the last year (in the past 12 months/year)
	The length of the rating scale (i.e. number of answer options) that the source variable uses	C_HEALTH_LENGTH	4 = 4-point source scale 5 = 5-point source scale 10 = 10-point source scale 11 = 11-point source scale
	Presence and position of the category 'fair' among response options in the source variable	C_HEALTH_FAIR	0 = source scale does not use response option 'fair' 1 = 'fair' is used as the midpoint of the source scale 2 = 'fair' is used as a response option for less than average health
	Source values: scale direction	C_HEALTH_ASCEND	0 = descending 1 = ascending

[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_HEALTH appear in 11 international survey projects: CB, EB, EQLS, ESS, EVS, ISSP, LB, LITS, NBB, NEB, WVS, 57 waves and 1221 national surveys. The data cover 117 countries and years from 1976 to 2017.

3. General Rules and Procedures

3.1. Source data description

To construct T_HEALTH_5 and T_HEALTH_DISTRIB we select source data files that include a measure of self-rated general health (the majority of cases) or a measure of self-rated physical health. The control variable C_HEALTH_PHYS indicates which type of question a given source data file provides.

Source variables of subjective health assessment use scales of varying lengths. Most frequent are 5-point scales (in 38 out of 48 data files), followed by 4-point scales (6 out of 48 data files). In three data files, self-assessed health is measured with a 10-point scale. In a single instance (EB/5 v84) health is measured with an 11-point scale. We show this variability in Appendix B, Table B.1 and capture it with the control variable C_HEALTH_LENGTH.

The source scales that respondents use to assess their health also vary in terms of direction. The majority of scales (in 37 out of the 48 source data files) rank health from best (value 1 = maximal value) to poorest health. In the remaining 11 data files (see Table 3.1.1), source scales start from worst health (1 = minimal value) to best health. We capture this variability with the control variable C_HEALTH_ASCEND.

Table 3.1.1. HEALTH. Types of scales.

Scale Length	Scale Direction	
	Descending	**Ascending**
4	EB/31A, 40, 52.1, 56.1, 62.2, 81.4	
5	EB/27, 34.1, 49, 58.2; EQLS/1-3; ESS/1-8; EVS/1-4; ISSP/2007, 2011, 2012, 2015; LB/2001, 2004-2007; LITS/1-3; NBB/1-6; NEB/5, 7; WVS/1-6	CB/2008-2015
10		EB/72.1, 74.1, 76.2
11		EB/5

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

For harmonization of T_HEALTH_5 and T_HEALTH_DISTRIB, we select source questions that use rating scales with four and more points. Projects with dichotomous scales, which we do not harmonize, are listed in Section 4.

If a source dataset contains both a measure of respondents' health in general, and direct (EB/27) or indirect measure(s) (EB/52.1, 56.1, 58.2, EQLS/1-3, ESS/3, 5-7, EVS/1-4, ISSP/2011, WVS/1-6) of their emotional/mental health, we select for harmonization the source variable that measures health in general (and assign value 0 on the control variable C_HEALTH_TYPE).

If a source dataset provides no general health question (e.g. LB/2001, NBB/1-6, NEB/5 and 7), but contains a direct question about respondents' physical health (e.g. "In the last 12 months, would you say your physical health has been very good, good, about average, bad or very bad?," source variable s21 from LB/2007), we take it (with value 1 assigned to C_HEALTH_TYPE).

To construct 5-point and distribution-preserving target variables, first we create preparatory scales. To achieve this, we recode the values of the source scales using the consecutive numbers k, where k ranges from 1 to n. The value 1 of the preparatory scale corresponds to worst health, and higher

scores correspond to better health (ascending direction). Each preparatory scale is of the same length as the source scale it was derived from.

3.2.1 Constructing 5-point target scale

To construct the 5-point rating scale T_HEALTH_5, we assign scores in the interval from 0 to 4, according to the following linear transformation applied to the preparatory scales:

$$l(k) = \frac{4}{n*2} + (k - 1) * \frac{4}{n}$$

where l(k) is a target score corresponding to the preparatory score k, and n is the number of k-values. This process involves "stretching" preparatory (and thus, source) scales that have fewer than 5 points, and "shrinking" scales that have more than 5 answer options. Table 3.2.1 provides assigned scores for source scales that appear in the SDR2 database.

Table 3.2.1. Creating the 5-point scale (from 0 to 4), with median and mean values 2, and minimized inter-scale differences in the variability

Source scale length	Recodes	Median Mean	Average of absolute deviations	Variance	Standard deviation
11-point	0.18, 0.55, 0.91, 1.27, 1.64, 2.00, 2.36, 2.73, 3.09, 3.45, 3.82	2.0	0.99	1.45	1.21
10-point	0.20, 0.60, 1.00, 1.40, 1.80, 2.20, 2.60, 3.00, 3.40, 3.80	2.0	1	1.47	1.21
5-points	0.0, 1.0, 2.0, 3.0, 4.0	2.0	0.96	1.60	1.26
4-points	0.50, 1.50, 2.50, 3.50	2.0	1	1.67	1.29
3-point	0.67, 2.0, 3.33	2.0	0.89	1.18	1.09

3.2.1 Constructing distribution-preserving target scale

To construct T_HEALTH_DISTRIB, we take into account respondents' position in the distribution of self-assessed health in a given national survey. For an n-point preparatory scale (cf. Step 1), for values k, k = 1,…,n, where Xi is the percent distribution of the variable in sample s, k is recoded to:

$$k = \sum_{i=1}^{k-1} X_i + \frac{X_k}{2}$$

For a given sample, the scale points of T_HEALTH_DISTRIB correspond to the mid-points of the cumulative distribution of scores k (see Table 3.2.2). Put differently, the scores of T_HEALTH_DISTRIB are percentiles that indicate what share of respondents within a national sample reports the same or poorer health than the individual. The target variable is computed using unweighted samples.

Table 3.2 illustrates how we transform **preparatory** variables (which recode **source** variables' values in ascending direction) with 5 response options into the distribution-based target variable.

Table 3.2. Example of the distribution-based transformation of 5-point preparatory variables into T_HEALTH_DISTRIB.

Preparatory variable values, based on source values k	**Percentage distribution** X_k	**Cumulative percentage distribution** $\sum_{i=1}^{k} X_i$	**Interval** $\sum_{i=1}^{k-1} X_i$	**Interval lower bound plus interval midpoint** $\sum_{i=1}^{k-1} X_i + \frac{X_k}{2}$	**Target value (rounded to integer)**
1 = worst health	10.68	10.68	0	= 10.68/2 = 5.34	5
2	32.75	43.44	10.68	= (10.68 + 32.75)/2 = 27.05	27
3	32.11	75.55	43.44	= (43.44 + 32.11)/2 = 59.49	59
4	21.69	97.23	75.55	= (75.55 + 21.69)/2 = 86.39	86
5 = best health	2.77	100	97.23	= (97.23 + 2.77)/2 = 98.61	99

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_HEALTH_5 and T_HEALTH_DISTRIB

We provide five harmonization control variables that capture specific features of the "health assessment" source variables (see Table 1.1):

C_HEALTH_PHYS is dichotomous, showing that the source question specifically asks about respondents' physical health (value 1), or their health in general (value 0).

C_HEALTH_TIME is a nominal variable that indicates whether respondents' state of health is assessed without reference to any specific time frame (value 0), with reference to the time around the interview (value 1), or with reference to the past year (value 2)

C_HEALTH_LENGTH is a nominal variable that identifies the length (i.e. number of answer options) of any given source scale used to construct T_HEALTH_5 and T_HEALTH_DISTRIB.

C_HEALTH_FAIR is a nominal variable. The control takes the value 0 if "fair" does not appear among a scale's answer options. It takes the value 1 if "fair" constitutes the midpoint of odd-numbered source scales (i.e. "fair" corresponds to average health). When "fair" appears as an answer option that should capture below-average health (e.g., in a 5-point scale with response options: excellent (1); very good (2); good (3); fair (4); worst (5)), we assign C_HEALTH_FAIR the value 2.

C_HEALTH_ASCEND is dichotomous. It takes the value 1 for source scales whose values are in ascending order (i.e. response options are ordered from worst to best health); it takes the value 0

when source scale values are in descending order (i.e. response options are ordered from best to worst health).

4. Special Cases

- EB/3 and EB/21 - we did not use the subjective health assessment questions from these projects as both of them are accompanied by dichotomous scales.
- EQLS/1-3, ISSP/2011, and EB/58 - we omitted the questions about the presence of an illness found in these projects.
- EQLS - we used two source variables, one for wave 1 and another one for waves 2 and 3.
- ISSP/2012 and ISSP/2015 - both datasets have C_HEALTH_FAIR set to 2 (meaning that 'fair' is used as a response option for less than average health), however the documentation indicates that Hungary used "bad" in ISSP/2012 and "poor" in ISSP/2015 instead of "fair".

Appendix A: 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 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.