HOUSEHOLD INCOME - General Target Variable Report (GVR)

1. General Information

The target variable T_INCOME_DISTRIB measures the relative position of a respondent in the distribution of household (HH) incomes in a given source survey (national sample). The scores of the target variable are percentiles within the national sample that indicate what share of participants reports the same or lower household income than the respondent. Percentile values of T_INCOME_DISTRIB are ascending, from minimum to maximum.

T_INCOME_DISTRIB does not adjust for the number of HH members or consumer units. Such adjustment was not feasible because a large number of source surveys ask about estimated deciles or other relative measures, or provide broad pre-coded brackets of income that are not suitable for computing an approximation of „individual income." We suggest that T_INCOME_DISTRIB be used together with the target variable T_HH SIZE (number of people in the household) as a control variable.

T_INCOME_DISTRIB is accompanied by two harmonization control variables, C_INCOME_TIME and C_INCOME_NOINCOME (see Table 1.1). C_INCOME_TIME indicates the timeframe for which the household income is reported (e.g., HH income for the last month/monthly HH income). C_INCOME_NOINCOME flags source variables that include the "no income" category in answer options (see Section 3.3 for details).

The target variable report for **Household Income** is accompanied by the following Excel documents:

- The Detailed Variable Report (DVR): T_INCOME_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_INCOME_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

Table 1.1. HOUSEHOLD INCOME: Description of the target, source, and control variables

	Variable description	**Variable name**	**Variable values**
Target variable	Household income (distribution-preserving scale)	T_INCOME_DISTRIB	0 = lowest percentile point in distribution 100 = highest percentile point in distribution

Source variables			See: T_INCOME_DVR_SDR2.xlsx T_INCOME_CWT_SDR2.xlsx
Control variables	Timeframe for respondent's household income referenced in the source variable	C_INCOME_TIME	0 = source question asks about monthly household income/household income for the last month 1 = source question asks about yearly household income/household income for the last year 2 = source question uses other timeframe than last month/ year or it provides the choice between several reporting periods 3 = source question does not reference any timeframe
	Presence of "no income" category in source variable	C_INCOME_NOINCOME	0 = "No income" category is not present in answer options 1 = "No income" category is presented in answer options

[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_INCOME_DISTRIB appear in 21 international survey projects: ABS, AFB, AMB, ARB, ASES, CB, CDCEE, CNEP, EB, EQLS, ESS, EVS, ISJP, ISSP, LITS, NBB, NEB, PA1, PA2, PPE7N, WVS, 131 waves and 2253 national surveys. The data cover 142 countries and years from 1966 to 2017.

3. General Rules and Procedures

3.1. Source data description

To construct the target variable T_INCOME_DISTRIB, we use source items that explicitly ask respondents about their total household income, where income can be before or after taxes, expressed as a specific amount or as an income bracket, and can stem from various sources, such as salary, pensions, etc. We do not harmonize source items about individual, husband's/spouse's or parents' income. Neither do we harmonize questions that pertain to subjective assessment related to the sufficiency of household income.

If a source datafile provides more than one total household income measure, we always prefer income brackets variables to variables recording specific amounts. At the same time, we give preference to country-specific source variables over variables that data producers harmonized ex-ante (for exceptions see Section 4).

When some national surveys of a given survey wave provide country-specific information on income, but other surveys of the same wave provide data through the pre-harmonized income variable, we use information from both country-specific and pre-harmonized source variables (e.g. ISSP 1995).

Most often, source questions ask respondents about "household income," but terms like "family income" also occur. The examples of a typical question wording are "What is your gross household income, before tax or other deductions, from all sources? Please include any pensions and allowances, and income from interest or dividends." (ISSP 2005 AU_INC, Australia) or "Which category best represents your total family income from all sources in 1991 - before taxes or other deductions?" (ISSP 1992 usa116, USA) For deviations from standard question wording, see Section 4.

In general, source questions ask respondents about their total monthly household income or household income for the last month. However, there is substantial variation from this formulation, as some questions refer to income earned during the last year, during other periods, or do not specify any timeframe. We capture this variation via the harmonization control C_INCOME_TIME (see Table 1.1. and Section 3.3.)

Sometimes, source variables provide "No income" as an answer option. We flag these instances with the harmonization control C_INCOME_NOINCOME.

3.2. Rules of transformation of source variables into target variable

To construct the target variable, we proceed in two steps. First, we recode source variables so that their lowest value always corresponds to the lowest income, and their highest value to the highest income.

We treat "No income" as the lowest income category whenever it is present among answer options of a source question. If "no income" is combined with other missing codes, like "Don't know" or "No answer" (e.g. in PPE7N_IN V299 - "No income, D.K., N.A."), we assign it the SDR2 missing code *unfit* (see Table A1 in the Appendix).

Second, to construct T_INCOME_DISTRIB, we take into account respondents' position in the distribution of all household incomes in a given source survey (national sample). Source values, in ascending order (from the lowest to the highest household income) are assigned values of the mid-point from the cumulative distribution. For the source *n*-point scale, for values *k* ranging from 1 to *n*, where X_k is the distribution of the variable, *k* is recoded to:

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

This transformation produces a variable that contains information about the relative position of respondent's household income in the distribution of household incomes in a given survey (national sample). Put differently, the scores of T_INCOME_DISTRIB are percentiles that indicate what share of participants within a national sample reports the same or lower household income than the respondent. 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_INCOME_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 = lowest household income	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 = highest household income	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_INCOME_DISTRIB

The target variable T_INCOME_DISTRIB is accompanied by two harmonization control variables C_INCOME_TIME and C_INCOME_NOINCOME (see Table 1.1).

C_INCOME_TIME is a nominal variable that indicates the timeframe for which the household income is reported. Value 0 indicates that respondents were asked about their monthly total household income/ hh income for the last month. Value 1 indicates source questions on the yearly household income/ hh income for the last year. Value 2 indicates that source questions use other timeframes than last month/last year, or they provide the choice between several reporting periods. Value 3 means that the source variable does not specify a timeframe for household income.

C_INCOME_NOINCOME flags source variables that include the "no income" category in answer options (value 1). It takes the value 0 if the "no income" category is not present in a source variable. In SDR2 source questions that do not include the "no income" category in answer options are more common.

4. Special Cases

1. AFB_2 q90 – we took this variable although it does not ask specifically about household or family income, but about respondent+spouse income. The question wording is: "Before taxes, how much money do you (and your spouse together) earn per month?"
2. ASES V0275 - we treat "490,000 won or less" as the minimal category, assuming there was a typo and it actually means "49,000 won or less". In this way the answer options look consistent and go in ascending order.
3. ASES V0278 – we treat the source value 1 "No fixed source of regular income" as "No income" minimal category.
4. CNEP_3_HU provides two variables – Z.Hu.HHIncome1, which is a question on actual amount of the household income Z.Hu.HHIncome2, which offers income categories in brackets for those respondents who refused to report an exact amount. We take only CNEP_3_HU Z.Hu.HHIncome1for harmonization.
5. ISJP_1_2 provides three variables on the household income: a question on actual amount of household income v151 and two follow-up questions with income categories in brackets, in case respondents refused to report actual income – v152 and v153. We take only ISJP_1_2 v151 for harmonization.
6. CNEP_3_MX L.Income variable has a discrepancy between source labels in different documentation sources. According to the Spanish questionnaire and data dictionary, the variable is about household income, but according to the English questionnaire, it is about personal income. We decided to take this variable because we prioritize information from data dictionary; also because other

country-specific CNEP_3 L.Income variables mostly provide information about household income.

7. LITS_3 q223 – the original maximum value for household income in Bosnia and Herzegovina is 2147483647 convertible marks. This value seems suspiciously high. We marked it as an error and took the next highest value, 70000000, as the maximum.
8. Some source variables had undocumented source values which resembled missing codes. In these instances, we coded those undefined source values with the SDR2 missing code *err* (cf. SDR2 missing codes schema, see Appendix) and set the maximum at the highest value that is documented (labeled). Specifically:

    - NEB_3 s6bul (Bulgaria) – has no proper source documentation for the source value 9999. We code 9999 as *err* and set the value 3000 (in local currency) as the maximum for this variable.
    - NEB_4 s10rom (Romania) – we code source values 17 and 21 as *err* and set the source value 16 as the maximum.
    - NEB_4 s10sve (Slovenia) – we code source values 996, 998, and 999 as *err* and set the source value 600 as maximum.
    - NEB_6 D8ABUL (Bulgaria) – we code source value 999 as *err* and set the source value 500 as maximum.
    - NEB_6 D8ASVK (Slovakia) - we code source values 88 and 99 as *err* and set the source value 27 as maximum.
    - PPE7N_YU V184 we code source values 0, 98 and 99 as *err* and set the source value 70 as maximum.
9. NEB_3 s6bel (Belarus) and NEB_3 s6ukr (Ukraine) are documented only via the general source questionnaire, common for all countries. However, unlike the other income variables in NEB_3, these two variables have “No personal income” as a response category. Since no other source documentation is available and we cannot check what exactly these variables asked about, in the SDR database we coded the household income variables corresponding to Belarus and Ukraine in NEB_3 as missings.
10. ARB_1 q716incomedeciles, EQLS_1_3 Y11_Income, ISSP_2010 CA_INC – these variables have source values with many decimals, so we rounded them to two decimal points for calculation purposes.

Appendix: Codes for missing values in SDR2

In SDR2 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.