SETTLEMENT SIZE – General Target Variable Report (GVR)

1. General Information

The target variable T_SETTLE_SIZE_DISTRIB measures the relative position of a respondent in the distribution of settlement sizes in a given source survey (national sample).[1] The scores of this target variable are percentiles within the national sample that indicate what share of respondents reports living in the same or smaller settlement size than the individual (see Table 1.1. and Section 3.3).

The target variable report for Settlement Size is accompanied by the SDR Crosswalk Table (CWT) Excel file: T_SETTLE_SIZE_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about how we map source values to target values.

Table 1.1. SETTLEMENT SIZE: Description of the Target Variable

	Variable description	**Variable name**	**Variable values** [a]
Target variable	Size of respondent's settlement (distribution-preserving measure)	T_SETTLE_SIZE_ DISTRIB	0 = Lowest percentile point in distribution 100 = Highest percentile point in distribution
Source variables			See: T_SETTLE_SIZE_CWT_SDR2.xlsx

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

2. Survey Projects

Source variables that we harmonized into T_SETTLE_SIZE_DISTRIB appear in all 23 international survey projects, 171 waves and their corresponding 208 source data files. Put differently, the target is available for 3167 national samples and 154 countries and the period 1966 – 2017.

3. General Rules and Procedures

3.1 Source data description

The main differences in source variables describing the respondent's settlement size can be summed up as follows:

- *Availability of source coding schemas harmonized ex-ante:* Some survey projects offer source variables that use only country-specific codes to describe respondents'

[1] The SDR2 project also provides the target variable T_RURAL, which indicates whether the respondent lives in a rural or non-rural place.

settlement size. Other projects provide additional variables with harmonized categories that can be compared between countries. For example, the World Value Survey provides two 'locality type' variables. The first variable has detailed country-specific categories (e.g. Japan – 1. *12 major large cities i.e., Tokyo, Osaka, etc.*; 2. *150,000 more residents in cities*; 3. *50,000 to 150,000 residents in cities*; 4. *Up to 50,000 residents in cities*; 5. *Rural districts*), while in the second variable this information is recoded into three categories (1. *rural/village*; 2. *small/medium town*; 3. *large town*). There are. also project waves with country-specific categories only for some countries

- *Response origin:* in some survey projects respondents answered the question on settlement size; in others, it is the interviewer/coordinator who coded the settlement size of respondents.
- *Number of response categories and type of information stored in response values*: source variables on settlement size are stored (1) in numeric form (locality size), (2) as a textual description, or (3) a mix of both. We use all three types of multi-categorical source variables to construct T_SETTLE_SIZE_DISTRIB:

  (1) Source variables that record settlement size as ordered numerical intervals, for example NEB_1 source variable `s2` where 1 = < *5000*; 2 = *5000-19,999*; 3 = *20,000-100,000*; 4 = *100,000 - 1,000,000* and 5 = > *1,000,000*;

  (2) Source variables that record place of residence as ordered textual qualifiers, for example ABS_2 source variable `ir14` where 1 = *capital or megacity*; 2 = *regional center or other major cities*; 3 = *small city or town*; 4 = *village or countryside*; 5 = *suburb* and 6 = *remote area*); and

  (3) Source variables that record place of residence as a combination of numbers and textual qualifiers, for example ASES source variable `v0357` where 1 = *big cities (Stockholm, Gothenburg, Malmo)*; 2 = *other cities more than 3, 000 inhabitants* and 3 = *villages less than 3, 000 inhabitants*.

3.2 Rules of Transformation of source variables into target variable

3.2.1 Selecting source variables

To decide which variable to harmonize into T_SETTLE_SIZE_DISTRIB when a source data file provides more than one variable describing respondent's settlement size, we apply the following rules:

1. **Prioritize country-specific variables.** When both pre-harmonized and country-specific measures of the respondent's settlement size are available, we choose the country-specific variable.
2. **Prioritize numeric over textual information**. If all source country-specific or pre-harmonized variables have the same number of answer categories, we choose the ones whose answer categories are numerical (e.g. "less than 5000 inhabitants") rather than textual (e.g. "rural", "village").

3. **Prioritize the source variable with more detailed information about the respondent's settlement size.** If a source dataset contains more than one pre-harmonized settlement size variable and no country-specific variables, we select for harmonization the variable with more detailed and clear information about how values can be ordered from the smallest to the largest settlement size.
4. **Prioritize variables that originate from respondents' self-assessment of their settlement size.** If a dataset provides several country-specific measures of respondents' settlement size, we select for harmonization the source variable that measures respondents' own assessments, rather than the variable that originates from survey administrators.

For source variables that we considered but did not harmonize into T_SETTLE_SIZE_DISTRIB, such as source items with only two valid options (e.g., ABS_1 level3, ARB_2 v13 and q13), we assign the SDR2 missing code "insuf." We present the SDR2 missing code schema in the Appendix.

We do not include source variables that have only two response options to create the distribution-preserving target scale for T_SETTLE_SIZE_DISTRIB.

3.2.2 Creating preparatory variables

After selecting the source variables, we create a preparatory variable. This involves recording the values of the source variables using the consecutive numbers *k*, where *k* ranges from 1 to *n*. The value 1 of the preparatory variable corresponds to the smallest settlement size (village), and higher scores correspond to larger settlement sizes (city; ascending direction). Each preparatory variable has the same number of values as the source variable it was derived from.

To recode values of source variables from smallest to larger settlement sizes, we follow the following rules:

1. Whenever possible, we place the SUBURBS option (or equivalents, e.g. "Semi-Urban" or "Peri-Urban" in AFB_6 URBRUR, "Conurbation" in EB_8 v137 (for Great Britain), "Outer metropolitan" in ISSP_2003 au_size):

> (a) before the corresponding CITY/TOWN option (e.g. ESS_1 domicil: "Suburbs or outskirts of big city" is placed before "A big city"; ISSP_1985 v119 [US]: "Suburb, medium city" is placed before "City, 50.000 - 250.000", "Suburb, large city" is placed before "City, greater than 250.000", while both blocks - middle and large cities - are also separated and ordered accordingly);
>
> (b) between cities/towns of different sizes (e.g. ISSP_1987 v84 (Great Britain): "Suburbs" is placed between "Small city/ town" and "Big city"); exceptions: ABS_2 ir14 ("Suburb" is placed between "Small city or town" and "Regional center or Other major cities" and "Capital or Megacity"), EB_54_1 v403 (Great Britain) ("Suburban" and "Urban" are on the same level before "Metropolitan", etc), ISSP_1991 v122

(Netherlands) ("Suburbs" and "Cities 5000 - 50000" are placed on the same level before "Cities >= 100000");

(c) between RURAL and URBAN options (e.g. ASES V0353: "Suburban" is placed between "Rural" and "Urban").

2. If an option is ambiguous, we follow the original order as proposed in the source data dictionary (e.g. CNEP_4_ZA Geotype: "Rural Formal" is placed before "Tribal Area", which is before "Urban Informal", and before "Urban Formal"). In rare cases when the available source documentation does not allow us to understand a given option, we tag it with the SDR2 missing code "unfit" (e.g., "Smallholdings" or "Hostels" in ISSP_2003 za_size or “Other” option in ISSP_2002 v358).

3. If two (or more) options share the same descriptions or are very similar, we place them on the same level (e.g., ISSP_1992 s128: "Rural1" and "Rural2" are at the same level and before "Urban1" and "Urban2," which we also treat as one level). The most difficult cases are PA1 v297 (Netherlands and United States), PA2 v1284 (Netherlands), and PPE7N_NL V453, where some options are merged according to some common words used in descriptions (see T_SETTLE_SIZE_CWT_SDR2.xlsx for more details).

4. If named cities or regions are used in the source scale, we place them according to their population sizes (e.g. ISSP_1992 s128: "Stockholm" is placed before "Goteborgs region," which is before "Malmo etc. region"). An exception are regions in Israel in some ISSP files, because these regions correspond rather to administrative regions than urban/rural areas.

5. If the source data provide unnamed regions but also population sizes, we order options according to the sizes (e.g., ISSP_2005 DO_SIZE). If other descriptives allow for the proper ordering, we also make use of them (e.g. EB_42 v529 (Belgium): "Autre localite" is placed before "Localite secondaire," which is before "Localite urbaine," itself before "Grand centre").

3.2.3 Constructing the distribution-preserving target variable

To construct the distribution-preserving target variable, we take into account respondents’ position in the distribution of reported settlement size values in a given national sample. For an *n*-point preparatory scale, for values *k* that range from 1 to *n,* where X_i is the percent distribution of the variable in a sample, *k* is recoded to:

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

The distributional score for the answer option *k* is the sum of percentiles of all previous answer options up to *k-1* plus half of the percentile of the answer option *k*.

For a given sample, each scale point of the distribution target scale corresponds to the midpoint of the cumulative distribution of scores *k* (see Table 3.2). Put differently, the scores of the distributional target scale are percentiles that indicate what share of respondents within a national sample reports the same or lower value than the individual. The target variable is computed using unweighted samples.

Note that we do not include source variables that have only two response options to create the distribution-preserving target scale.

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_SETTLE_SIZE_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 = the smallest settlement size	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 = the largest settlement size	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.

4. Special cases

- ISJP_1_2 v11008: source option 10 lacks the label. Although it is not documented, we recode it on the basis of the order of other response options and additional distribution analysis.

- ISSP_1985 v119 (United States): We ordered source values based on their textual labels. As a result, our order is different from source order of this variable and it reads: *Open country < Smaller areas < Town, greater than 2.500 < City, 10.000 - 49.999 < Uninc, medium city < Suburb, medium city < City, 50.000 - 250.000 < Uninc, large city < Suburb, large city < City, greater than 250.000*.
- ISSP_1987 v84 (Austria): After additional analysis, we decided to combine some answer categories into one. Specifically, we ordered answer options as: *Village in rural area < Village in urban area < Town with industry + Town in rural area < Middle-sized city with large industrial zone + Middle-sized city without large industry < Metropolitan area-suburb < Metropolitan centre*
- ISSP_2003 za_size, ISSP_2004 v377, ISSP_2007 ZA_SIZE, ISSP_2008 ZA_SIZE: We coded “Smallholdings" and “Hostels” options with the SDR2 missing code “Unfit” as it was not clear from source documentation where to place this category in the sequence of other response options.
- Source variables PA1 v297 (Netherlands and United States), PA2 v1284 (Netherlands), and WVS variable X049CS for South Korea were difficult to order from the smallest to largest settlement size. Based on our analysis of source documentation and distributions, we made assumptions as to what settlement size the source options correspond to and proposed the mapping of source values to target values based on these observations. For details please consult T_SETTLE_SIZE_CWT_SDR2.xlsx.

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.