INTEREST IN POLITICS- General Target Variables Report (GVR)

1. General Information

The target variable T_INTPOL measures respondents' level of interest in (national) politics and public affairs. Source variables measuring the concept 'interest in politics' are harmonized into two versions of the target variable: T_INTPOL_5 and T_INTPOL_DISTRIB. T_INTPOL_5 is a result of rescaling into the 0-4 range. It takes the value 0 when respondents report that they are not interested in politics at all. It takes the value 4 when respondents indicate that they are very interested in politics. T_INTPOL_DISTRIB captures the relative position of an individual in the distribution of a given society with 0 indicating the lowest and 100 indicating the highest point in distribution (see Table 1.1).

The target variables are accompanied by the set of harmonization control variables specifying: 1) features of source question wording (C_INTPOL_NONSTANDARD and C_INTPOL_LEVEL); 2) properties of source scales, such as scale length (C_INTPOL_LENGTH) and scale direction (C_INTPOL_ASCEND).

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

- The Detailed Variable Report (DVR) T_INTPOL_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_INTPOL_CWT_SDR2.xlsx. CWT Excel files in SDR2 contain details about mapping of source values to target values.

Table 1.1. INTEREST IN POLITICS: Description of the target, source, and control variables

	Variable description	**Variable name**	**Variable values***
Target variable	Interest in politics (5-point scale)	T_INTPOL_5	0 = Lowest degree 4 = Highest degree
	Interest in politics (distribution-preserving scale)	T_INTPOL_DISTRIB	0 = Lowest percentile point in distribution 100 = Highest percentile point in distribution

Source variables			See: T_INTPOL_DVR_SDR2.xlsx T_INTPOL_CWT_SDR2.xlsx
Control variables	Source non-standard question wording	C_INTPOL_NONSTANDARD	0 = Standard question wording (general interest in politics) 1 = Interest in national / public affairs 2 = Interest in politics and affairs of the nation 3 = Interest in politics and government 4 = Interest in politics and economics 5 = Interest in economics and political affairs 6 = Interest and activity (participation) 7 = How often follow government and public affairs: frequency 8 = No source question wording available
	Source interest in politics scale length	C_INTPOL_LENGTH	3 = 3-point scale 4 = 4-point scale 5 = 5-point scale
	Source interest in politics scale direction	C_INTPOL_ASCEND	0 = Descending 1 = Ascending
	Question on interest in politics in general or on the national level	C_INTPOL_LEVEL	0 = General interest (no politics level specified) 1 = National, selected from set of questions 2 = National, only one question available

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

2. Survey Projects

Source variables that we used in harmonization of T_INTPOL appear in 19 international survey projects: ABS, AFB, AMB, ARB, ASES, CB, CNEP, EB, ESS, EVS, ISSP, LB, NBB, NEB, PA1, PA2, PPE7N, VPCPCE, WVS, 78 waves and 1480 national surveys. The data cover 155 countries and years from 1966 to 2017.

3. General Rules and Procedures

3.1. Source data description

To construct T_INTPOL, we use source items about respondents' self-reported interest in (national) politics and public affairs at the time of the survey. We rely on the English language and Spanish language questionnaires and codebooks describing the source survey data.

A typical question in the majority of the surveys is 'How interested would you say you are in politics?' (ABS).

Next to the standard item formulation discussed above, some source surveys use a different question wording to collect data on respondents' interest in politics. Specifically, respondents were asked how interested are they in *domestic/national/ public affairs* (AFB/2-6, EB/88 and PPE7N Nigeria), *politics and affairs of the nation* (PPE7N India, Japan, United States), *politics and government* (AFB/1 some countries), *politics and economics* (EB/60), *economic and political affairs* (PPE7N Yugoslavia), and *how often [do you] follow government and public affairs* (frequency not intensity question occur in some countries in AFB/1). There is also one case when a source question wording is not available in the master questionnaire (NEB/2).

3.2. Rules of transformation of source variables into target variable

First, to construct the 5-point and distribution-preserving target variables, 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 lower interest in politics, and higher scores correspond to higher interest (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 target scale, we use preparatory scales and assign scores to them in the interval from 0 to 4, according to the following linear transformation:

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

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 the distribution-preserving target scale, we take into account respondents' position in the distribution of reported values in the 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 sample *s*, *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 a sum of percentiles of all previous answer options (up to *k-1*) plus the half of the percentile of the answer option *k*.

For a given sample, each scale point of a distribution target scale corresponds to the midpoint of the cumulative distribution of scores *k* (see Table 3.2.2). 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.

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_INTPOL_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 interest in politics	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 interest in politics	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_INTPOL

The target variables are accompanied by harmonization control variables (see Table 1.1 and Section 3.3).

C_INTPOL_NONSTANDARD captures non-standard question wording. It takes the value 0 when the question wording is standard, e.g., *Generally speaking, how interested would you say you are in politics?* (ARB/1); it takes the value 1 when the question wording mentions interest in *national or public affairs* (e.g., AFB); it takes the value 2 when the question asks about interest in *politics and affairs of the nation* (e.g., PPE7N India); it takes the value 3 when the question asks about *interest in politics and government* (e.g., AFB/1 Mali); it takes the value 4 when the question asks about interest in *politics and economics* (e.g., EB/60.1); it takes the value 5 when the question asks about interest in *economics and political affairs* (e.g., PPE7N Yugoslavia); it takes the value 6 when the question asks about *interest* and *activity* (participation) (e.g., EB/0.1); it takes the value 7 when the question asked *how often follow government and public affairs* (frequency) (AFB/1 Botswana); and the value 8 when no source question wording available (e.g., NEB/2).

C_INTPOL_LEVEL identifies questions on interest in *national* politics versus questions on interest in general (level not specified). If the question asks about general interest in politics (e.g., ESS), we assign the value 0. If the question asks about interest in *national* politics and was selected from a set of questions (e.g., PPE7N Austria), we assign the value 1. If the question asks about interest in *national* politics and is the only question available (e.g., PPE7N India), we assign the value 2 to C_INTPOL_LEVEL.

C_INTPOL_LENGTH specifies the source scale length. It takes the value 3 whenever the source question contains 3 response options, e.g., 'How interested are you in public affairs?': 'not interested'; 'somewhat interested'; 'very interested' (AFB/2). It takes the value 4 when respondents are provided with 4 response options, e.g., 'How interested are you in politics?': 'very interested'; 'fairly interested'; 'a little interested'; 'not at all interested' (LB/2010). It takes the value 5 if respondents can choose between 5 response options, e.g., 'How interested would you say you personally are in politics?': 'fairly interested'; 'somewhat interested'; 'not very interested'; 'not at all interested' (ISSP/2006).

C_INTPOL_ASCEND identifies scale direction. It is coded as 1 if the source response scale is ascending, i.e., responses are ordered from the least to the most interested: 'How interested would you say you are in politics?': 'not very interested'; 'somewhat interested'; 'very interested' (ABS/1). Otherwise, it is coded as 0, e.g., 'To what extent would you say you are interested in politics?': 'a great deal'; 'to some extent'; 'not much', 'not at all' (EB/42).

4. Special cases

1. In EB/0.1 the question is "DO YOU PARTICIPATE IN POLITICAL ACTIVITIES, OR DO YOU FOLLOW POLITICS WITH INTEREST WITHOUT ACTIVELY PARTICIPATING, OR DO POLITICS NOT INTEREST YOU MORE THAN OTHER THINGS, OR NOT AT ALL?" with the response categories:
   "1|PARTICIPATE PERS 2|INTEREST NO PARTICIP 3|ONLY SLIGHT INTEREST 4|NO INTEREST AT ALL". We have marked 1 = the strongest interest, 4 = no interest. The nonstandard question wording is captured by the control variable.
2. In AFB/1 the question was about interests in "government and public affairs": "Some people seem to follow what's going on in government and public affairs most of the time, whether there's an election going on or not. Others aren't that interested. Would you say you follow what's going on in government and public affairs:_____?". Moreover, the question differed by countries: "Question text - GHA/MALI/NIG/TNZ/UGA How interested are you in politics and government?". The response categories for the two types of questions differed as well. For the first case they were: "Always/Most of the time; Some of the time; Only now and then; Hardly at all; Don't know (DNR) " and in the second variant of questions: "Very interested, Somewhat interested, Not interested, Don't know (DNR)" with the note that ""Don't know" was not offered as a response

option in Ghana." - In the dataset we coded control variables on scale length and question wording (nonstandard) at the country level for this datafile.

3. In PPE7N Netherlands: the question asked not about interest, but about attention to national matters: "DO YOU PAY A LOT OF ATTENTION TO NATIONAL MATTERS, SOME ATTENTION, OR NO ATTENTION AT ALL?". We did not take this question to create a target variable.
4. PPE7N United States codebook suggests a four-point scale: 1. NOT INTERESTED 2. ONLY SLIGHTLY INTERESTED 3. SOMEWHAT INTERESTED 4. VERY INTERESTED 9. DK 0. NA, while realized values use only three points: 0 1 2 3 9. It is unclear if the values are rescaled in data, so that they start from 0 and end with 3, or if 0 captures "don't know", as the codebook suggests. As we have only a codebook available, we code this variable as a four point scale, where 0 is mapped to a target "Not elsewhere classified" and source values 1, 2 and 3 are rescaled following harmonization rules for four point scales.
5. In PPE7N India: response categories in the codebook have an order from 3 = very interested to 0 = not interested, which suggests a descending scale direction (from very interested to not interested). However, we do not know if the scale presented to a respondent in the questionnaire was not ordered as numeric values suggest, ascending from 0 = not interested to 3 = very interested. Since the questionnaire is not available, we decided to code the control variable C_INTPOL_ASCEND as 0 = descending, but to rescale the target as if it was ascending.
6. PPE7N India scale has undefined value 7: (3) VERY INTERESTED (2) SOMEWHAT INTERESTED (0) NOT INTERESTED (7) (8) OTHER (9) D.K., N.A. As we have no reasons to assume that the value 7 was a part of the scale, we code the scale length control variable C_INTPOL_LENGTH as = 3. 7 is coded as a type of missing data ERR (-9 / .i).
7. CB/2009 asks a set of questions about interest in POLICY rather than interest in politics. We did not include questions on interest in 'policy'.
8. EB/0.2, 3, 13, 40 ask about problems in the "European community" only. We did not include them in the set of source variables used to create these target variables.
9. CNEP/3 and 4 recodes are available, but we selected source questions from the questionnaires when possible. Also in CNEP/3 Mexico the question on interest in politics was asked before elections and repeated after (for the same sample). Here and elsewhere we select the responses from the PRE-election period.
10. In WVS/2 Czech Republic and Slovakia - out of a four point scale in the codebook only two options were used ("realized" in data): "very interested" and "somewhat interested". We use this variable for the distributional scale target T_INTPOL_DISTRIB. However we do not use this source when constructing the five point target variable because of the possible bias in the mean value. Source values for these two countries in the T_INTPOL_5 are coded as UNFIT (-8 .h).

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.