Harmonized Income Dataset version 1.0

Documentation

January 29, 2019

Authors: Marta Kołczyńska, Przemek Powałko

Acknowledgements: We thank Ilona Wysmułek and Denys Lavryk for their assistance.

Funding: This work was supported by the grant "Effects of status inconsistency on political values, attitudes and behavior: a cross-national analysis with survey data harmonized *ex post*" funded by the Institute of Philosophy and Sociology, Polish Academy of Science (2018), and the 2017 Silverman Research Support Award from the Department of Sociology, The Ohio State University.

Table of contents

Table of contents 2
1. General information 3
2. Technical variables 5
3. Household income 6
3.1. Availability of source variables 6
3.2. General rules and procedures 7
3.2.1. Source data description 7
3.2.2. Rules of transformation of source variables into target variables 10
3.2.3. Harmonization control variables 11
3.3. Special Cases 12
4. Personal income 13
4.1. Availability of source variables 14
4.2. General rules and procedures 15
4.2.1. Source data description 15
4.2.2. Rules of the transformation of source variables into target variables 16
4.2.3. Harmonization control variables 17
4.3. Special Cases 18
5. Codebook 20
Technical variables 20
Personal income 25
Household income 24
6. References 26

1. General information

The Harmonized Income Dataset provides harmonized individual-level survey variables on personal and household income from major cross-national survey projects, as well as technical variables necessary to match them to the Survey Data Recycling version 1 dataset (SDR v.1, Slomczynski et al. 2017), which contains harmonized survey items on political participation, political attitudes, as well as their selected correlates.

The SDR v.1 dataset contains data from 22 cross-national survey projects according to the following criteria: (1) non-commercial character, (2) cross-national, and – ideally – multi-wave design, (3) samples intended as representative of adult populations, (4) presence of questions on political participation and attitudes, and (5) availability of documentation in English (SDHT 2017: 6). Altogether, the SDR v.1 dataset includes data on 2,289,060 respondents in 1721 national surveys from 142 countries/territories surveyed between 1966 and 2013. Table 1.1 describes the survey projects in terms of time span, number of waves, national surveys, and cases, i.e., respondents (SDHT 2017: 7).

Table 1.1. Projects included in SDR v.1.

Abbrev.	Project name	Time	Number of		
			waves	national surveys	cases
ABS	Asian Barometer	2001-2011	3	30	43691
AFB	Afrobarometer	1999-2009	4	66	98942
AMB	Americas Barometer	2004-2012	5	92	151341
ARB	Arab Barometer	2006-2011	2	16	19684
ASES	Asia Europe Survey	2000	1	18	18253
CB	Caucasus Barometer	2009-2012	4	12	24621
CDCEE	Consolidation of Democracy in Central and Eastern Europe	1990-2001	2	27	28926
CNEP	Comparative National Elections Project[1]	2004-2006	1	8	13372
EB	Eurobarometer[3]	1983-2012	7	152	138753
EQLS	European Quality of Life Survey	2003-2012	3	93	105527
ESS	European Social Survey	2002-2013	6	146	281496
EVS	European Values Study	1981-2009	4	128	166502
ISJP	International Social Justice Project	1991-1996	2	21	25805
ISSP	International Social Survey Programme[3]	1985-2013	13	363	493243
LB	Latinobarometro	1995-2010	15	260	294965
LITS	Life in Transition Survey	2006-2010	2	64	67866
NBB	New Baltic Barometer	1993-2004	6	18	21601
PA2	Political Action II[2]	1979-1981	1	3	4057
PA8NS	Political Action - An Eight Nation Study	1973-1976	1	8	12588
PPE7N	Political Participation and Equality in Seven Nations	1966-1971	1	7	16522
VPCPCE	Values and Political Change in Postcommunist Europe	1993	1	5	4723
WVS	World Values Survey	1981-2008	5	184	256582
	Total	1966-2013	89	1721	2289060

[1] Only post-election samples; [2] Only cross-sectional samples. [3] Selected waves.

The SDR v.1. dataset also includes information on methodological characteristics of the included surveys, such as their sampling type, appropriate information content of the survey documentation, processing errors, duplicated. A separate file includes selected macro-level indicators with political, economic, and demographic characteristics of countries.

We reviewed the source datasets and documentation from all projects included in the SDR v.1 dataset for availability of variables on personal and household income (Table 1.1). Of the 22 international survey projects we reviewed, source variables measuring household income are available in 1177 national surveys in 18 projects, and individual income in 453 national surveys in 10 projects. 419 national surveys in 9 projects have both variables. Table 1.2 shows the availability of both income items in national surveys for each analyzed survey project.

Altogether the dataset includes 1,464,035 cases (records) in 1211 national surveys, from 65 project waves in 19 projects, covering 135 countries between 1966 and 2013.

Table 1.2. Availability of income items in projects included in SDR v.1.

Abbrev.	Project name	N national surveys			
		Total	with household income	with personal income	with both
ABS	Asian Barometer	30	30	0	0
AFB	Afrobarometer	66	15	0	0
AMB	Americas Barometer	92	90	0	0
ARB	Arab Barometer	16	16	16	16
ASES	Asia Europe Survey	18	18	0	0
CB	Caucasus Barometer	12	12	12	12
CDCEE	Consolidation of Democracy in Central and Eastern Europe	27	0	27	0
CNEP	Comparative National Elections Project[1]	8	3	3	1
EB	Eurobarometer[3]	152	53	0	0
EQLS	European Quality of Life Survey	93	93	0	0
ESS	European Social Survey	146	134	0	0
EVS	European Values Study	128	128	0	0
ISJP	International Social Justice Project	21	20	20	20
ISSP	International Social Survey Programme[3]	363	356	355	350
LB	Latinobarometro	260	0	0	0
LITS	Life in Transition Survey	64	0	0	0
NBB	New Baltic Barometer	18	18	9	9
PA2	Political Action II[2]	3	3	3	3
PA8NS	Political Action - An Eight Nation Study	8	8	6	6
PPE7N	Political Participation and Equality in Seven Nations	7	4	2	2
VPCPCE	Values and Political Change in Postcommunist Europe	5	0	0	0
WVS	World Values Survey	184	176	0	0
	Total	1721	1177	453	419

2. Technical variables

Technical variables are necessary to match the Harmonized Income Dataset to the SDR v.1 dataset. These variables are: the abbreviation of the source data file (T_SOURCE_TABLE), the case ID (T_CASE_ID), the survey project name (T_SURVEY_NAME), the survey project edition (T_SURVEY_EDITION), the country code (T_COUNTRY_L1U), and the country set (T_COUNTRY_SET). The combination of these variables uniquely identify a record of the data. Additionally, for convenience, two variables were added: the year in which the survey was carried out (T_COUNTRY_YEAR) and the calibrated case weight (T_WEIGHT_L1U_2).

Detailed information about these technical variables are available in the "Master File Documentation. SDR Master Box Version 1.0" (SDHT 2017).

3. Household income

This section describes a set of harmonized target variables pertaining to reported *household income*, named T_INCOME_HH, T_INCOME_HH_RANK_100, T_INCOME_HH_PROP_100, and T_INCOME_HH_DISTRIB[1]. These target variables capture the within-survey variation in respondents' household income, mapped in four different ways. The target variable T_INCOME_HH contains household income as declared by the respondent and recorded in the original units of the source variables; the target T_INCOME_HH_RANK_100 maps household income to the continuous rank scale from 0 to 100, where 0 is no income and 100 is the highest point in the distribution; the target T_INCOME_HH_PROP_100 offers a mapping of household income values to the proportional scale, where 0 means the lowest income and 100 means the highest income, and distances between values correspond to the distances on the original scales; and T_INCOME_HH_DISTRIB maps household income into distribution-preserving scale, where 0 is the lowest point in distribution and 100 is the highest point in the distribution, and values correspond to the proportion of the sample with income below the given respondent in the same national sample.

These four target variables were constructed on the basis of the relevant source variables (S_INCOME_HH) referring to the respondent's household income. Source variables on respondents' personal income, sometimes available in survey projects together with the declared household income, were not taken into account for construction of this set of targets (for harmonized personal income see Section 4 of this document).

The set of target variables is accompanied with the control variable C_INCOME_HH_SCALE_LENGTH, capturing the number of distinct values represented in the data, ranging from 4 to 622.

The four target variables allow for the within-survey ranking of individuals with regard to household income, with slight differences in how the ranking is done depending on the target variable. It is worth emphasizing, that these target variables do not allow comparisons between national surveys to describe, for example, the variation in average household income across countries.

Table 3.1 presents the variable names, labels, and value codes and labels.

3.1. Availability of source variables

The source variables on respondent's household income used to create a set of target variables appeared in 1177 national surveys within 63 project waves of 18 international survey projects covering 134 countries at various points from 1966 till 2013. The file: HID_details_v1.xlsx provides detailed information on data coverage.

[1] This document followed a modified version of the template for documenting harmonized target variables described in Wysmułek et al. (2015).

Table 3.1. Description of target variables on HOUSEHOLD INCOME.

	Variable label	Variable name	Variable values*
Target variables	Household income (raw value)	T_INCOME_HH	0 = no income 90 000 000 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Household income (continuous 0-100 rank scale)	T_INCOME_HH_RANK_100	0 = lowest income 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Household income (proportional 0-100 scale)	T_INCOME_HH_PROP_100	0 = lowest income 100 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Household income (distribution-preserving scale)	T_INCOME_HH_DISTRIB	0 = lowest point in distribution 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
Source variables	Source value of personal income	S_INCOME_HH	See: HID_details_v1.xlsx Not included in the public file.
Control variable	Source household income scale length	C_INCOME_HH_SCALE_LENGTH	4 = "4" 622 = "622" -2 (.b) = not applicable

* Missing value codes: SPSS (Stata).

3.2. General rules and procedures

3.2.1. Source data description

Individual economic status is a necessary element of almost all sociological analyses, including studies of political attitudes and behavior. Of the different possible survey variables measuring economic status, including individual or household earnings, individual or household income, wealth, and perceived economic standing, household income items are the most popular: they are available in 1177 national surveys out of the 1721 gathered in the SDR v.1.0 project (cf. Table 1.1 for the list of projects included in SDR v.1, and Table 1.2 for the availability of income items across survey projects).

Although the basic concept – household income - is the same in all 1177 national surveys, the way of operationalizing and measuring it differs substantially across surveys with regard to the formulation of the question, and the way of recording responses. First, surveys ask about net income (after taxes and transfers) or gross income, and about weekly, monthly, or annual income. These differences often occur within projects across national surveys, so controlling for the question wording would require the examination of country questionnaires in local languages, which are often not available. However, this does not hurt the within-survey ranking of people according to household income, which is the primary aim of this harmonization effort.

Second, in some cases the exact amount is recorded (usually in the local currency), but in others it is the income category or quantile (decile or quintile). Additionally, the surveys differed in how the values were coded in the source data file: in some surveys income categories were assigned mid-points of the respective intervals, while other surveys coded income with consecutive ordinal numbers.

Source variables measuring respondent's income vary in the formulation of the question and the way of coding answers not only between waves and projects, but also within waves, including the codes for missing values. An additional complication is the presence of country-specific variables (separate variables for each country in the wave dataset). As a result of all these differences, the recoding of source into target values had to be done for each national survey separately, making it a long and laborious process that required numerous quality and consistency checks.

For example, in the Asia Europe Survey, household income is represented with separate variables for each country. For Japan, responses are recorded in categories in Japanese Yen, and coded from 1 (lowest income) to 12 (highest income), with 31 reserved for “no answer”. In the same project in South Korea, responses are coded from 1 (lowest income) to 8 (highest income), while 9 corresponds to “refused”, and 31 means “no answer”.

Another example of within-wave differences comes from the Arab Barometer, Wave 1, where all countries except for Morocco have household income recorded as the exact value in the local currency (coded from 0 to 15000000), while Morocco coded income in categories from 0 (no income) to 9 (highest income), with missing value codes of 10 (“not exact”) and 100 (“not provided/not usable”).

The first three rounds of the European Social Survey measured income with categories in the local currency, giving respondents a choice between weekly, monthly, and annual income. Starting with Round 4, ESS constructs response categories to correspond to the deciles of the income distribution in the given country. In many ESS national surveys the resulting distribution of household income does not resemble a uniform distribution, which questions the meaning of the income categories as deciles. The project Political Participation and Equality in Seven Nations on the other hand, asked about monthly income in India, the Netherlands, and Yugoslavia, but about annual income in Japan and Nigeria, in each case recording income categories of equal or varying size.

The variation among household income items in the World Values Survey is described in Donnelly and Pop-Eleches (2016). Although WVS records income on a 10-point scale, the meaning of the categories combine (a) income brackets resembling deciles of the country distribution and determined by the country's PI, (b) questions about exact household income that are then recoded into deciles, and (c) questions about subjective perception of respondent's income relative to others in the country on a 1-10 scale. As Donnelly and Pop-Eleches (2016) show, these differences result in variation in the meaning and country distributions of the resulting variables.

As a result, the target variables on household income were constructed on the basis of a very diverse set of source variables.

The fact that some source variables recorded household income in categories, some in quantiles, and some asked about the actual precise value, affects the within-survey distributions of the raw values of household income: from left-skewed, to bimodal, to uniform, to almost normal, to heavily right-skewed. Table 3.2 contains the distribution of kurtosis and skewness of the within-survey distributions.

The analyzed surveys also have varying amounts of item non-response in the household income variables, which is a common problem in survey research. Given the distribution of the proportion of missing values, which is continuous without obvious gaps, we decided to eliminate only two extreme cases: ISSP/2007/DE-E and ISSP/2007/DE-W, both of which have a single non-missing case. In both cases we abstained from using this source variable for constructing target variables, so these two surveys do not have harmonized target variables for household income.

The next survey with the highest amount of missing values is ABS/2/CN, which only has valid responses for 0.11 of the sample, i.e. 89% of the cases have missing values (compared to the mean of 17% missing cases across all analyzed surveys). Despite this, we include ABS/2/CN in the Harmonized Income Dataset.

The characteristics of survey distributions of original household income variables is presented in Table 3.2. Information about the proportion of missing values and the length of the original scale for each national survey is available in the file `HID_details_v1.xlsx`.

Table 3.2. Characteristics of the source household income variables.

Variable (*N* = 1177 surveys)	**Min**	**Q1**	**Median**	**Mean**	**Q3**	**Max**
Kurtosis (weighted)	-1.803	-0.770	0.031	22.713	4.556	1097.242
Skewness (weighted)	-1.411	0.077	0.618	1.858	1.728	33.047
Proportion missing	0.000	0.077	0.140	0.170	0.240	0.889

3.2.2. Rules of transformation of source variables into target variables

The methodological variation described in Section 3.2.1 makes it hardly possible to harmonize household income in terms of assigning each respondent a monetary value in some common metric. This would require data, for every country and year, about the income distribution in the population (to convert quantiles into quantities), and the purchasing power of the local currency (to make measurements comparable). While the latter is in principle feasible, data about income distributions are not readily available.

We chose an alternative solution to the harmonization problem that satisfies our needs regarding household income used primarily as a control variable. We propose to map source variables on household income into target variables that preserve the within-survey rank order of the respondents with regard to household income (`T_INCOME_HH_RANK_100`), and capture the relative position of the individual in the survey distribution (`T_INCOME_HH_PROP_100` and `T_INCOME_HH_DISTRIB`), but do not allow to compare means of household income across samples.

The harmonization process consisted of the following steps:

1. Identification of candidate source variables for harmonization in the source data files for 22 survey projects (see Table 1.1 for the survey project list).
2. Investigation of the candidate variables to eliminate unharmonizable cases due to, e.g., the unknown meaning of income categories in ISJP/1996/EE (Wegener and ISJP 2010: 49).
3. If more than one source variable is available: selection of the best source variable for harmonization (i.e., the one with fewest missing values and/or largest number of response options).
4. Identification of missing value codes in household income variables for each national survey.
5. Recoding variables coded in the descending order (only in CB and CNEP/3/ES) into the ascending order by reversing the coding.

The resulting, cleaned variable is stored as `T_INCOME_HH`.

6. Assigning the target value 0 to lowest realized values.
7. Assigning ranks (consecutive integers) to source values.
8. Rescaling to 0-100, where 0 correspondents to “lowest income” and 100 corresponds to “highest income”.

The resulting variable is stored as `T_INCOME_HH_RANK_100`. This variable has equal intervals between each pair of consecutive realized values.

Starting with the `T_INCOME_HH` variable:

9. Assigning the target value 0 to the lowest realized values,
10. Rescaling to 0-100 while preserving the relative distances between realized values.

The resulting variable is stored as `T_INCOME_HH_PROP_100`.

Starting with the T_INCOME_HH variable:

11. Assigning the position (mid-point of relevant category) from the cumulative distribution. The resulting variable is stored as T_INCOME_HH_DISTRIB.
12. Creating control variables for the number of distinct values realized in each national survey.

Table 3.3. shows the recoding example of substantive values, i.e. after missing values are identified and assigned appropriate missing value codes, from the source variable (S_INCOME_HH) to target variables (T_INCOME_HH_*) on the example of Afrobarometer, Round 3, Mozambique.

3.2.3. Harmonization control variables

The set of target variables for household income is accompanied by one harmonization control variable: C_INCOME_HH_SCALE_LENGTH, which records the number of distinct values represented in the source variable (in the data, not in the codebook). These distinct values may correspond to either income categories or exact values, depending on the design of the original question.

The values of the C_INCOME_HH_SCALE_LENGTH variable range from 4 (EVS/4/MD, EVS/4/ME, WVS/2/KR) to 622 (ISJP/1991/NL). Most surveys have few response categories: the median value is 11, and the third quartile is 21. Out of the total 1177, there are 209 surveys that have 100 or more values of the original household income variable represented in the data.

Table 3.3. Transformation of household income variable to a set of harmonized variables for the Afrobarometer survey from Mozambique in wave 2 (AFB/2/MZ).

S_INCOME_HH	*N*	T_INCOME_HH	T_INCOME_HH_RANK_100	T_INCOME_HH_PROP_100
0	697	0	0	0
3	340	3	16.67000008	30
4	167	4	33.33000183	40
6	100	6	50	60
8	43	8	66.66999817	80
9	19	9	83.33000183	90
10	12	10	100	100
Total	1378			

S_INCOME_HH	*N*	Relative distribution (%)	Cumulative distribution (%)	T_INCOME_HH_DISTRIB
0	697	50.58	50.58	25
3	340	24.67	75.25	63
4	167	12.12	87.37	81
6	100	7.26	94.63	91
8	43	3.12	97.75	96
9	19	1.38	99.13	98
10	12	0.87	100	100
Total	1378	100		

3.3. Special Cases

ISJP/1996/EE: not harmonizable due to the unknown meaning of income categories (cf. Wegener and ISJP 2010: 49).

CNEP/3/MX: According to the English questionnaire the question is about personal income; according to the Spanish questionnaire it is household income. Since much of the harmonization work relies on English language documentation and documentation in other language (if available) is consulted only in special cases, in this case the variable is coded as personal income.

If "no income" and "don't know" or some other non-response category shared the same code (the case of **ISSP/2004/PH**), the target variable code corresponds to the non-response category ("don't know").

> "in Philippines (PH):
>
> 999997 Refused, no comment, the saved money is used for the expenses
>
> 999998 Don't know, can't estimate, no fixed income"
>
> (ISSP Research Group 2012: 242)

In **CB (all surveys) and CNEP/3/ES** the source values of the household income variables are coded in the descending order, i.e. the highest code value is assigned to the lowest income category. In these cases the original coding is reversed, i.e. the original coding scheme:

7 (lowest income), 6, 5, 4, 3, 2, 1 (highest income)

is transformed into the scale:

1 (lowest income), 2, 3, 4, 5, 6, 7 (highest income).

4. Personal income

This section describes a set of harmonized target variables pertaining to reported *personal income*: T_INCOME_PERSONAL, T_INCOME_PERSONAL_RANK_100, T_INCOME_PERSONAL_PROP_100, and T_INCOME_PERSONAL_DISTRIB. These personal income variables were harmonized in an analogous way to household income described in Section 3, but on the basis of different source variables. The personal income target variables capture the within-survey variation in respondents' personal income, mapped in four different ways. Similar to household income, the target variable T_INCOME_PERSONAL contains income as declared by the respondent in the original units, as recorded in the source variables; the target T_INCOME_PERSONAL_RANK_100 maps income to the continuous rank scale from 0 to 100, where 0 is no income and 100 is the highest point in the distribution; the target T_INCOME_PERSONAL_PROP_100 offers a mapping of income values to the proportional scale, where 0 means the lowest income and 100 means the highest income, and distances between values correspond to the distances on the original scales; and T_INCOME_PERSONAL_DISTRIB maps income into distribution-preserving scale, where 0 is the lowest point in distribution and 100 is the highest point in the distribution, and values correspond to the proportion of the sample with income below the given respondent in the same national sample.

These four target variables were constructed on the basis of the relevant source variables (S_INCOME_PERSONAL) referring to the respondent's household income. Source variables on respondents' household income, sometimes available in survey projects together with the declared personal income, were not taken into account for construction of this set of targets (for harmonized household income see Section 3 of this document).

The set of target variables is accompanied with the control variable C_INCOME_PERSONAL_SCALE_LENGTH, capturing the number of distinct values represented in the data, ranging from 6 to 928.

The four target variables allow for the within-survey ranking of individuals with regard to household income, with slight differences in how the ranking is done depending on the target variable. It is worth emphasizing that these target variables do not allow comparisons between national surveys to describe, for example, the variation in average household income across countries.

Table 4.1 presents the variable names, labels, and value codes and labels.

Table 4.1. Description of the target variables on PERSONAL INCOME.

	Variable label	Variable name	Variable values*
Target variables	Personal income (raw value)	T_INCOME_PERSONAL	0 = no income 75 000 000 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Personal income (continuous 0-100 rank scale)	T_INCOME_PERSONAL_RANK_100	0 = lowest income 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Personal income (proportional 0-100 scale)	T_INCOME_PERSONAL_PROP_100	0 = lowest income 100 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
	Personal income (distribution-preserving scale)	T_INCOME_PERSONAL_DISTRIB	0 = lowest point in distribution 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
Source variables	Source value of personal income	S_INCOME_PERSONAL	See: HID_details_v1.xlsx Not included in the public file.
Control variable	Source personal income scale length	C_INCOME_PERSONAL_SCALE_LENGTH	6 = "6" 928 = "928" -2 (.b) = not applicable

* Missing value codes: SPSS (Stata).

4.1. Availability of source variables

Of the 1721 national surveys we reviewed, source variables with information about respondent's personal income appeared in 453 national surveys within 30 project waves of 10 international survey projects covering 71 countries at various points from 1966 till 2013. The file HID_details_v1.xlsx provides detailed information on data coverage.

4.2. General rules and procedures

4.2.1. Source data description

Personal income was harmonized in addition to the target variables on household income (described in Section 3) on the basis of source variables capturing Respondent's personal (individual) income. Items on personal income are much less popular in cross-national surveys than questions about household income: the former appear in 453 out of the 1721 gathered in the SDR v.1.0 project (cf. Table 1.1 for the list of projects included in SDR v.1, and Table 1.2 for the availability of income items across survey projects).

Similarly to the case of household income, the way of measuring personal income differs substantially across surveys with regard to the formulation of the question, and the way of recording responses. First, surveys ask about net income (after taxes and transfers) or gross income, and mention weekly, monthly, or annual income. As in the case of household income, we did not record these differences in form of control variables, because they frequently occur within projects across national surveys, while country questionnaires in local languages for many projects are not available. However, the formulation of the question does not substantially affect the within-survey ranking of people according to household income, which is the primary aim of this harmonization effort.

Second, in some cases the exact amount is recorded (usually in the local currency), but in others it is the income category or quantile (decile or quintile). Additionally, the surveys differed in how the values were coded in the source data file: in some surveys income categories were assigned mid-points of the respective intervals, while other surveys coded income categories with consecutive ordinal numbers.

Source variables measuring respondent's income vary in the formulation of the question and the way of coding answers not only between waves and projects, but also often within waves, including with regard to the codes for missing values. An additional complication is the presence of country-specific variables (separate variables for each country in the wave dataset), as well as the coding of "no income" as a separate code or together with lowest income category. As a result of all these differences, the recoding of source into target values had to be done for each national survey separately, making it a long and laborious process that required numerous quality and consistency checks.

For example, in ISSP/1995, personal income is represented with a single variable for all countries, but the coding ranges differ across countries: 0-10 in West Germany and Australia 1-90. Most other waves of ISSP has country-specific variables, coded in categories and sometimes with exact values recorded (e.g., ISSP/2006/FI).

Another example of within-wave differences comes from the Arab Barometer, Wave 1, where all countries except for Morocco have household income recorded as the exact value in the local

currency (coded from 0 to 30,000,000), while Morocco coded income in categories from 0 (no income) to 9 (highest income), while the value 10 corresponds to "not exact".

The project Political Participation and Equality in Seven Nations only asks about personal income in India and Japan, and about monthly income in India but about annual income in Japan, with a different number of categories.

The fact that some source variables recorded personal income in categories, some in quantiles, and some the asked about the actual precise value, affects the within-survey distributions of the source values of household income: from left-skewed, to bimodal, to uniform, to almost normal, to heavily right-skewed. Table 4.2 contains information about the kurtosis and skewness of the within-survey distributions of the source variable.

The analyzed surveys also have varying amounts of missing data, ranging from no missing values in several surveys to 80% missing values in ARB/2/DZ. A summary of the distribution of the characteristics of survey distributions of personal income is presented in Table 4.2. Information about the proportion of missing values and the length of the original scale for each national survey is available in the file `HID_details_v1.xlsx`.

Table 4.2. Characteristics of the harmonized target variable T_INCOME_PERSONAL.

Variable (N = 453 surveys)	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Kurtosis	-1.651	1.045	6.997	41.927	27.842	1322.008
Skewness	-0.648	1.018	2.001	3.387	3.923	34.762
Proportion missing	0.000	0.050	0.009	0.139	0.193	0.799

4.2.2. Rules of the transformation of source variables into target variables

The methodological variation described in Section 4.2.1 makes it hardly possible to harmonize personal income in terms of assigning each respondent a monetary value in some common metric. This would require the knowledge, for every country and year, about the income distribution in the population (to convert quantiles into quantities), and the purchasing power of the local currency (to make measurements comparable). While the latter is in principle feasible, data about income distributions are not readily available.

We chose an alternative solution to the harmonization problem that satisfies our needs regarding personal income used primarily as a control variable. The constructed target variables preserve the rank order of the respondents with regard to household income (`T_INCOME_PERSONAL_RANK_100`), and capture the relative position of the individual in the survey distribution (`T_INCOME_PERSONAL_PROP_100` and `T_INCOME_PERSONAL_DISTRIB`), but do not allow to compare means of household income across samples.

The harmonization process consisted of the following steps:

1. Identification of candidate variables for harmonization in the source data files,
2. If more than one source variable is available: selection of best source variable for harmonization (i.e., the one with fewest missing values and/or largest number of response options),
3. Identification of missing value codes for each national survey,
4. If necessary, recoding of variables coded in the descending order (CB, CNEP/3/ES) into the ascending order by reversing the coding,

The resulting, cleaned variable is stored as `T_INCOME_PERSONAL`.

5. Assigning the target value 0 to lowest realized values,
6. Assigning ranks (consecutive integers) to source values,
7. Rescaling to 0-100, where 0 correspondents to "lowest income" and 100 corresponds to "highest income".

The resulting variable is stored as `T_INCOME_PERSONAL_RANK_100`. This variable has equal intervals between each pair of consecutive realized values.

Starting with the `T_INCOME_PERSONAL` variable:

8. Assigning the target value 0 to lowest realized values,
9. Rescaling to 0-100 while preserving the relative distances between realized values.

The resulting variable is stored as `T_INCOME_PERSONAL_PROP_100`.

Starting with the `T_INCOME_PERSONAL` variable:

10. Assigning of the position (mid-point of relevant category) from the cumulative distribution.

The resulting variable is stored as `T_INCOME_PERSONAL_DISTRIB`.

11. Creating control variables for the number of distinct values realized in each national survey.

A recoding example of substantive values, i.e. after missing values are identified and assigned appropriate missing value codes, from the source variable (`S_INCOME_HH`) to target variables (`T_INCOME_HH_*`) on the example of ISSP/1998/Ireland, is presented in Table 4.3.

4.2.3. Harmonization control variables

The target variable "Household income" is accompanied by one harmonization control variable: `C_INCOME_PERSONAL_SCALE_LENGTH`, which records the number of distinct values represented in the source variable (in the data, not in the codebook). As a reminder, these distinct values may correspond to either income categories or exact values, depending on the design of the original question.

The values of the `C_INCOME_ PERSONAL _SCALE_LENGTH` variable range from 6 (CNEP/3/ES, ISSP/1998/PT, ISSP/2004/PT, ISSP/2008/PT, ISSP/2009/PT, and PPE7N/JP) to 928

(ISSP/1990/AU). Most surveys have few response categories: the median value is 21, and the third quartile is 120.

Table 4.3. Transformation of household income variable to a set of harmonized variables for International Social Survey Project, wave 1998, Ireland (ISSP/1998/IE).

S_INCOME_PERSONAL	*N*	T_INCOME_PERSONAL	T_INCOME_PERSONAL_RANK_100	T_INCOME_PERSONAL_PROP_100
0	245	0	0	0
25	69	25	11.11	3.57
75	198	75	22.22	10.71
125	118	125	33.33	17.86
175	90	175	44.44	25
250	164	250	55.56	35.71
350	70	350	66.67	50
450	38	450	77.78	64.29
600	11	600	88.89	85.71
700	7	700	100	100
Total	1010			

S_INCOME_PERSONAL	*N*	Relative distribution	Cumulative distribution	T_INCOME_HH_DISTRIB
0	245	24.26	24.26	12
25	69	6.83	31.09	28
75	198	19.60	50.69	41
125	118	11.68	62.38	57
175	90	8.91	71.29	67
250	164	16.24	87.52	79
350	70	6.93	94.46	91
450	38	3.76	98.22	96
600	11	1.09	99.31	99
700	7	0.69	100	100
Total	1010	100		

4.3. Special Cases

CNEP/3/MX: According to the English questionnaire the question is about personal income; according to the Spanish questionnaire it is household income. Since much of the harmonization work relies on English language documentation and documentation in other language (if available) is consulted only in special cases, in this case the variable is coded as personal income.

If "no income" and "don't know" or some other non-response category shared the same code (as in **ISSP/2004/PH**), the target variable code corresponds to the non-response category.

> "in Philippines (PH):
> 999997 Refused, no comment, the saved money is used for the expenses
> 999998 Don't know, can't estimate, no fixed income"
> (ISSP Research Group 2012: 238)

In the **CB** and **CNEP3/ES** the source values of the personal income variables are coded in the descending order, i.e. the highest code value is assigned to the lowest income category. In these cases the original coding is reversed, i.e. recoded as follows:

7 (lowest income), 6, 5, 4, 3, 2, 1 (highest income)

is transformed into the scale:

1 (lowest income), 2, 3, 4, 5, 6, 7 (highest income).

PA8NS/USA and Finland, had income coded from 1 (no income) to 7 (highest income). Since in other cases “no income” was typically coded as 0, we subtracted 1 from the source values so that the target values range from 0 (no income) and 6 (highest income).

5. Codebook

Technical variables

Variable label	Variable name	Variable values*
T_SOURCE_TABLE	Internal source table name (corresponds to source data file)	ABS_1, ABS_2, ABS_3, AFB_2, AMB_1_5, ARB_1, ARB_2, ASES, CB_2009, CB_2010, CB_2011, CB_2012, CDCEE_1_2, CNEP_3_ES, CNEP_3_HU, CNEP_3_MX, CNEP_3_PT, CNEP_3_UY, EB_1983, EB_1984, EB_1989, EB_2000, EQLS_1_3, ESS_1_5, ESS_6, ISJP_1_2, ISSP_1985, ISSP_1989, ISSP_1990, ISSP_1991, ISSP_1996, ISSP_1998, ISSP_2004, ISSP_2006, ISSP_2007, ISSP_2008, ISSP_2009, ISSP_2010, ISSP_2011, IVS_1_9, NBB_1_6, PA2, PA8NS, PPE7N_IN, PPE7N_JP, PPE7N_NG, PPE7N_NL
T_CASE_ID	Unique case identifier within source table/data file	[case number in the source data file]
T_SURVEY_NAME	Project name	ABS = Asian Barometer AFB = Afrobarometer AMB = Americas Barometer ARB = Arab Barometer ASES = Asia Europe Survey CB = Caucasus Barometer CDCEE = Consolidation of Democracy in Central and Eastern Europe CNEP = Comparative National Elections Project EB = Eurobarometer EQLS = European Quality of Life Survey ESS = European Social Survey EVS = European Values Study ISJP = International Social Justice Project ISSP = International Social Survey Programme NBB = New Baltic Barometer PA2 = Political Action II PA8NS = Political Action = An Eight Nation Study PPE7N = Political Participation and Equality in Seven Nations WVS = World Values Survey
T_SURVEY_EDITION	Project edition (wave/round)	[project-specific number of the project wave; *null* for standalone projects]
T_COUNTRY_YEAR	Year in which national survey was conducted	1966, 1968, 1971, 1973, 1974, 1975, 1976, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1989, 1990, 1991, 1992, 1993, 1994, 1995, 1996, 1997, 1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013
T_COUNTRY_L1U	Country/territory code, level 1 unit	AD = Andorra AL = Albania AM = Armenia AR = Argentina AT = Austria AU = Australia AZ = Azerbaijan BA = Bosnia and Herzegovina BA-FBH = Federation of Bosnia and Herzegovina BA-RSR = Republika Srpska

BD = Bangladesh
BE = Belgium
BE-FLA = Belgium-Flanders
BE-WAL = Belgium-Wallonia
BF = Burkina Faso
BG = Bulgaria
BO = Bolivia, Plurinational State of
BR = Brazil
BW = Botswana
BY = Belarus
BZ = Belize
CA = Canada
CH = Switzerland
CL = Chile
CN = China
CO = Colombia
CR = Costa Rica
CV = Cape Verde
CY = Cyprus
CY-TCC = Turkish Cypriote Community
CZ = Czech Republic
DE = Germany
DE-E = East Germany
DE-W = West Germany
DK = Denmark
DO = Dominican Republic
DZ = Algeria
EC = Ecuador
EE = Estonia
EG = Egypt
ES = Spain
ET = Ethiopia
FI = Finland
FR = France
GB = United Kingdom
GB-GBN = Great Britain
GB-NIR = Northern Ireland
GE = Georgia
GH = Ghana
GR = Greece
GT = Guatemala
GY = Guyana
HK = Hong Kong
HN = Honduras
HR = Croatia
HT = Haiti
HU = Hungary
ID = Indonesia
IE = Ireland
IL = Israel
IL-ARB = Israel-Arabs
IL-JEW = Israel-Jews
IN = India
IQ = Iraq
IR = Iran, Islamic Republic of

IS = Iceland
IT = Italy
JM = Jamaica
JO = Jordan
JP = Japan
KE = Kenya
KG = Kyrgyzstan
KH = Cambodia
KR = Korea, Republic of
KS = Kosovo
LB = Lebanon
LS = Lesotho
LT = Lithuania
LU = Luxembourg
LV = Latvia
MA = Morocco
MD = Moldova, Republic of
ME = Montenegro
MK = Macedonia, the former Yugoslav Republic of
ML = Mali
MN = Mongolia
MT = Malta
MW = Malawi
MX = Mexico
MY = Malaysia
MZ = Mozambique
NA = Namibia
NG = Nigeria
NI = Nicaragua
NL = Netherlands
NO = Norway
NZ = New Zealand
PA = Panama
PE = Peru
PH = Philippines
PK = Pakistan
PL = Poland
PR = Puerto Rico
PS = Palestine, State of
PT = Portugal
PY = Paraguay
RO = Romania
RS = Serbia
RU = Russian Federation
RU-KRA = Krasnoyarsk
RW = Rwanda
SA = Saudi Arabia
SD = Sudan
SE = Sweden
SG = Singapore
SI = Slovenia
SK = Slovakia
SN = Senegal
SV = El Salvador
TH = Thailand

		TN = Tunisia TR = Turkey TT = Trinidad and Tobago TW = Taiwan, Province of China TZ = Tanzania, United Republic of UA = Ukraine UG = Uganda US = United States UY = Uruguay VE = Venezuela, Bolivarian Republic of VN = Viet Nam YE = Yemen ZA = South Africa ZM = Zambia ZW = Zimbabwe
T_COUNTRY_SET	Dataset version within wave	1 = dataset 1 2 = dataset 2
T_WEIGHT_L1U_2	Re-calibrated target value of weight	0 = 0 90.31986 = 90.31986

* Missing value codes: SPSS (Stata).

Household income

Target variables		
T_INCOME_HH	HOUSEHOLD INCOME (RAW VALUE)	0 = no income 90 000 000 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_HH_RANK_100	HOUSEHOLD INCOME (CONTINUOUS 0-100 RANK SCALE)	0 = lowest income 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_HH_PROP_100	HOUSEHOLD INCOME (PROPORTIONAL 0-100 SCALE)	0 = lowest income 100 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_HH_DISTRIB	HOUSEHOLD INCOME (DISTRIBUTION-PRESERVING SCALE)	0 = lowest point in distribution 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
Control variable		
C_INCOME_HH_SCALE_LENGTH	SOURCE HOUSEHOLD INCOME SCALE LENGTH	4 “4” 622 “622” -2 (.b) = not applicable
* Missing value codes: SPSS (Stata).		

Personal income

Target variables		
T_INCOME_PERSONAL	PERSONAL INCOME (RAW VALUE)	0 = no income 75 000 000 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_PERSONAL_RANK_100	PERSONAL INCOME (CONTINUOUS 0-100 RANK SCALE)	0 = lowest income 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_PERSONAL_PROP_100	PERSONAL INCOME (PROPORTIONAL 0-100 SCALE)	0 = lowest income 100 = highest income -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
T_INCOME_PERSONAL_DISTRIB	PERSONAL INCOME (DISTRIBUTION-PRESERVING SCALE)	0 = lowest point in distribution 100 = highest point in distribution -9 (.i) = missing data -8 (.h) = question not asked in national survey -5 (.e) = variable not identified in data file -1 (.a) = don't know
Control variable		
C_INCOME_PERSONAL_SCALE_LENGTH	SOURCE PERSONAL INCOME SCALE LENGTH	6 “6” 928 “928” -2 (.b) = not applicable
* Missing value codes: SPSS (Stata).		

6. References

Donnelly, Michael J., and Grigore Pop-Eleches. 2016. "Income Measures in Cross-National Surveys: Problems and Solutions." Political Science Research and Methods 6 (2): 355–63. https://doi.org/10.1017/psrm.2016.40.

ISSP Research Group. 2012. "ISSP 2004 – Citizenship I. Variable Report". GESIS-Variable Reports No. 2012/199. Documentation release 2012/06/27. Related to the international dataset Archiv-Study-No. ZA3950 Version 1.3.0, https://doi.org/10.4232/1.11372.

Slomczynski, Kazimierz M., J. Craig Jenkins, Irina Tomescu-Dubrow, Marta Kołczyńska, Ilona Wysmułek, Olena Oleksiyenko, Przemek Powałko, and Marcin W. Zieliński. 2017. "SDR Master Box." 2017. https://doi.org/10.7910/DVN/VWGF5Q.

SDHT (Survey Data Harmonization Team). 2017. "Masterfile documentation." SDR Master Box Version 1. https://doi.org/10.7910/DVN/VWGF5Q.

Słomczyński, Kazimierz M., Irina Tomescu-Dubrow, and J. Craig Jenkins. 2016. *Democratic Values and Protest Behavior. Harmonization of Data from International Survey Projects*. Warsaw: IFiS Publishers.

Wegener, Bernd, and International Social Justice Project (ISJP). International Social Justice Project, 1996 and 2000 [Germany]. ICPSR22750-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-02-08. http://doi.org/10.3886/ICPSR22750.v1ICPSR

Wysmułek, Ilona, Olena Oleksiyenko, Przemek Powałko, Marcin W. Zieliński, Kazimierz M. Słomczynski. 2015. "Towards Standardization: Target Variable Report Template in the Harmonization Project." Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences 1(2): 13-17. https://www.asc.ohio-state.edu/dataharmonization/wp-content/uploads/2015/11/harmonization-newsletter-v1n2-fall-2015-with-issn-_final-3.pdf (2019.01.29).