Ex-post Survey Data Harmonization Toolbox: A Reproducible and User-Friendly Worlflow*

Marta Kołczyńska

Institute of Philosophy and Sociology, Polish Academy of Sciences

30 March, 2019

Abstract

Ex-post harmonization of survey data creates new opportunities for research by extending the geographical and/or time coverage of analyses. While an increasing number of scholars combine different survey projects to analyze them as a single dataset, there is substantial variation in the understanding of what ex-post harmonization entails, and with regard to the methods of data processing, documentation, as well as the reproducibility of all procedures on the basis of materials the projects make publicly available. In this paper I propose a procedure and a set of simple tools for the exploration, recoding, and documentation of harmonization of survey data, relying on crosswalks and a combination of automation for improved reproducibility and efficiency, with human decision-making that allows for flexibility necessary in dealing with the variation and diverse standard observed in survey datasets. The resulting documentation of the harmonization process is user-friendly and readible without the knolwedge of any programming language. The presented tools rely on the programming language R and spreadsheets - both common software among social scientists. Harmonization of variables on trust in institutions from three cross-national survey projects serve as an illustrate of the proposed workflow.

*Corresponding address: kolczynska.1@osu.edu. Acknowledgments: I am grateful to all those generous people who invest their time into open-source software and reproducible research.

Contents

1 Introduction

With extensive data archives in place and further archiving efforts in progress, the focus on combining and merging of existing data in order to build on what had already been done seems to be the optimal strategy and logical next step of developing the research infrastructure, and has already received supportive recognition (Burkhauser & Lillard, 2005).

International, and even multi-wave, survey projects are a major advancement in the social science infrastructure for cross-national research, yet their country and time coverage remains necessarily limited by funding availability, organizational conditions, and PIs' interests. Ex-post harmonization of survey data promises to overcome these limitations to create larger datasets with more global coverage. At the same time, ex-post harmonization can create country time series necessary for longitudinal analyses, which enables stronger tests of theoretical mechanisms than cross-country comparisons. Thus, harmonization of existing data in order to maximally exploit their potential for research is one of the major challenges in the social sciences. While the promises of new research opportunities are alluring to many, the associated challenges are multi-faceted - including technical, logistical, methodological, and substantive - and not yet well understood. This paper addresses one of these challenges, concerning the organization and documentation of the ex-post survey data harmonization process in a way that enables reproducibility of all data processing, as well as the cumulative character of harmonization efforts.

The number of projects relying on combined data from different cross-national survey projects is increasing. These projects range from small initiatives where the multi-projects dataset is created for the purposes of a single publication or dissertation (e.g., Christmann, 2018; Christmann & Torcal, 2018; Mauk, 2019) to large projects performing ex-post survey data harmonization not - or not only - for own substantive research, but in order to improve the social science research infrastructure (Bekkers et al., 2015; Klassen, 2018; Slomczynski & Tomescu-Dubrow, 2018). Correspondingly, the effort invested in the creation and sharing of the documentation of the harmonization process is characterized by substantial variation.

The diversity of approaches to documenting the harmonization process reflects the lack of established and followed standards for computational reproducibility, i.e. the ability to re-create the results of published research using materials - data and code - provided by the authors of the original study (cf. Liu & Salganik, 2019). The value of published research largely depends on its reproducilibity (Buckheit & Donoho, 1995). In addition to data and code, researchers are starting to recognize the role of the programming environment in reproducibility of results both currently, and in the near and more distant future. For example, Liu & Salganik (2019) propose to use software containers (such as Docker) to standardize the software and cloud computing (such as Amazon Web Services) to standardize the hardware, especially in computation-intensive analyses.

Reproducibility not only concerns research publications, but also processed data, such as those resulting from ex-post survey data harmonization products, although in this case it might be more accurately called 'data processing reproduciblity', thus excluding data modeling.

In this paper I propose a procedure and a set of simple tools for the exploration, recoding, and documentation of harmonization of survey data, relying on crosswalks for mapping one coding scheme onto another. The proposed approach includes automated steps that ensure reproducibility and efficiency of data processing, with human decision-making to integrate methodological expertise and domain knowledge, and enable flexibility necessary in dealing with the variation and diverse standard observed in survey datasets. The product of the harmonization process is its documentation in form of crosswalk tables that map (1) source variables to target variables and (2) source values to target values. The human-friendly character and readibility of crosswalks enables the verification and reproduction of the harmonization process.

The example presented in this paper uses the programming language R and spreadsheets - both common software among social scientists. The general framework combining recoding syntax and spreadsheets is software-agnostic and can be used with any programming language. Survey variables on trust in institutions from three cross-national survey projects - the European Social Survey, the European Values Study, the European Quality of Life Survey - serve as an illustration of the proposed workflow.

This work builds on the experiences of four inter-related projects: (1) the *Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling* project funded by the (Polish) National Science Centre (2012/06/M/HS6/00322, 2013-2016), its continuation *Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research* funded by the U.S. National Science Foundation (2017-2021, PTE Federal award 1738502), (3) *New Approach to Analyses of the Relationship between Democracy and Trust: Comparing European Countries Using Quantitative and Qualitative Methodology* funded by the (Polish) National Science Centre (2012/05/N/HS6/03886, 2013-2015), (4) *Stratified modernity, trust in state institutions and democratic change in cross-national perspective: using harmonized survey data from 116 countries* project supported by the Silverman Research Support Award from the Department of Sociology, The Ohio State University (2017), and (5) *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 (2017/S/05, 2018).

The proposed procedures are substantially simplified compared to those implemented in the aforementioned projects in the spirit of accessibility and reproducibility, and the harmonization tools were developed independently. Altogether, the workflow proposed in this paper is sufficiently simple that it can be successfully implemented by a single person or a small team of programming non-specialists, and powerful enough that it can handle large amounts of data and harmonization situations of moderate complexity, with highly complex cases possible to accomodate after some modifications.

2 What is ex-post survey data harmonization?

Ex-post (or retrospective) data harmonization refers to procedures applied to already collected data to improve the comparability and inferential equivalence of measures collected by different studies (Fortier et al., 2017). In the case of ex-post survey data harmonization, the procedures are applied to survey data sets that were not intended for joint analysis, in order to integrate them into a single dataset that can be meaningfully employed in substantive analyses. The harmonization process is simultaneously theory-informed and data-driven. Theories provide the concepts to be measured, but data availability to a large extent determines what ends up being measured and how. The following steps can be distinguished (cf. Granda & Blasczyk, 2016; Fortier et al., 2017; Kołczyńska, 2019; Wolf, Schneider, Behr, & Joye, 2016):

(1) concept definition:

  a. defining the concept of interest guided by the theoretical framework and hypotheses;
  b. based on this definition, developing a preliminary coding scheme or choosing a coding scale for the harmonized (target) variable;

(2) data preparation:

  a. gathering available surveys (data and documentation) that meet the requirements regarding the topics, target population and representativeness, and potentially others;
  b. exploring the methodological variation they represent with regard to the design of the survey items of interest and the overall survey process;
  c. describing surveys in terms of methodology, including quality (and potentially exclude some surveys on this basis), and constructing survey quality indicators;
  d. identifying relevant items in the gathered surveys that correspond to the target concept(s);
  e. describing the methodological variation in the design of the selected survey items given the research of survey methodology and effects of item design on respondents' answers (potentially exclude some items in this step);
  f. identifying relevant dimensions of variation between the survey items to be captured by harmonization control variables;
  g. adjusting the coding scheme or scale of the harmonized (target variable) based on the observed variation in the survey items;

(3) harmonization:

   a. transforming source variables into the target variable(s) using the coding scheme;
   b. constructing harmonization control variables to capture the properties of source variables that would be lost in the process of recoding;

(4) checking the target variable for errors and documentation of the whole process.

2.1 Ex-post survey data harmonization projects: A non-exhaustive review

Researchers increasingly combine data from multiple survey projects to fill in the gaps in geographical coverage, or to increase the presence of typically underrepresented countries, e.g. autocracies or countries from the Global South. Identifying projects that turn to ex-post survey data harmonization is not easy, because the term “ex-post harmonization” is not often used. This section briefly describe the approaches taken in a non-representative sample of ex-post survey data harmonization projects. The number of publications relying on survey data harmonized ex-post is difficult to know, because the term “ex-post survey data harmonization” is not widely used as the name of this procedure. An alternative term, “integrative data analysis”, proposed by Curran and Hussong (2009), has not caught on either.

Based on a non-systematic review of ex-post survey data harmonization projects, two broad categories can be distinguished: small-scale projects where harmonization is performed for the purposes of a single paper or dissertation, and large-scale projects that enable multiple different analyses with the general aim of improving the social science research infrastructure.

2.1.1 Small-scale projects

The two small-scale projects I identified, ex-post survey data harmonization provided data for dissertation research. The first project (Christmann, 2018; Christmann & Torcal, 2018) harmonized data from several cross-national and national survey projects. The selection criterion for the projects was a particular design of the satisfaction with democracy question, which was the key variable in the analysis: only surveys where a four-point response scale was used were harmonized. Replication materials that accompany the paper, available from Harvard's Dataverse, include the final harmonized data file and code for conducting the analysis, but no harmonization syntax[1].

The second project deals with political trust and democratic values, and relies on data from six cross-national survey projects from a hundred countries and almost global coverage (Mauk, 2019). Replication materials, also stored on Dataverse, include a recoding syntax file with all the code necessary to create the harmonized variables used in subsequent analyses. The harmonized dataset is not provided[2].

2.1.2 Infrastructure projects

There are at least three large-scale projects that perform ex-post survey data harmonization. The first one is the Democratic Values and Protest Behavior (DVPB) project (Slomczynski & Tomescu-Dubrow, 2018, p. @Slomczynskietal2016), which harmonized selected variables on political participation, political trust, and basic socio-demographics, from 22 major international survey projects between 1966 and 2013. The data are publicly available via Harvard's Dataverse[3] The project, now completed, used a suite of open-source tools for extracting source data, applying the harmonization procedures, and outputting data files in a format usable by social scientists. The programming environment is based on scripting tools, while data processing occurs in a MySQL database (Powałko & Kołczyńska, 2016). The publicly available documentation consists of codebooks, recode syntax in SQL and lists of source variable names selected for harmonization from each dataset (Slomczynski et al., 2017). The continuation of this initiative, the Survey Data Recycling project (dataharmonization.org), will extend the set of harmonized variable to

[1] Christmann, Pablo; Torcal, Mariano, 2017, "Replication Data for: The Effects of Government System Fractionalization on Satisfaction with Democracy", https://doi.org/10.7910/DVN/EU541C, Harvard Dataverse, V1.

[2] Mauk, Marlene, 2019, "Replication Data for: Disentangling an elusive relationship", https://doi.org/10.7910/DVN/IIVAJM, Harvard Dataverse, V1.

[3] The data and documentation from the Democratic Values and Protest Behavior (DVPB) project are available via Dataverse at https://doi.org/10.7910/DVN/VWGF5Q

also include social capital and well-being indicators, and update the data up until 2017.

The second project, Human Understanding Measured Across National Surveys (HUMANS, humansurveys.org) uses data from 19 cross-national and national survey projects, many of them the same as in the previous project, and harmonized variables on social trust, satisfaction with democracy, support for democracy, and perceived electoral integrity, as well as basic socio-demographics. The data are also freely available.[4] In terms of documentation, the codebook (Klassen, 2018) provides the recodes from source to target coding schemes, names of source variables, and names of source data files, all in a PDF document.

The third initiative, the Harmonized Trust Database created by the Global Trust Research Consortium (GTRC) (globaltrustresearch.wordpress.com), contains data from 79 national and cross-national survey projects, covering 155 countries since 1953. The project Open Science Framework profile[5] provides a PDF codebook (Sandberg & Bekkers, 2018) with names of source projects, but without data file versions or source variable names. The harmonized data are not publicly available.

The three projects have very similar goals and scopes - in terms of the type of source data and the substantive interest in harmonizing particular variables - and yet, their approaches to documenting the harmonization, and the extent the harmonization is reproducible, are very different. While the DVPB project published voluminous documentation, in practice the exact replication of all harmonization procedures would most likely not be straightforward, despite the natural language-resembling nature of SQL, due to complications with re-creating the software environment. The HUMANS project published source and target variables, but the recoding schema is provided in a PDF document - not the most machine-readable format. GTRC provides neither the recode schemas, nor the resulting data.

The three large project described in this section have two other things in common. First, with the data and documentation made public by these projects it is not possible to extend the

[4] The data and documentation from the Human Understanding Measured Across National Surveys (HUMANS, humansurveys.org) projects are available at https://dataverse.harvard.edu/dataverse/humansurveys.

[5] The Harmonized Trust Database created by the Global Trust Research Consortium (GTRC) (globaltrustresearch.wordpress.com) makes the documentation available via OSF at https://osf.io/qfv76/

dataset by adding extra harmonized target variables. The reason is that the two projects that provide the harmonized data - DVPB and HUMANS - do not include original respondent (case) IDs from the source data files, so even though it is possible to construct additional variables separately, it is impossible to match them to the already harmonized data. Second, while all three are performing very similar and labor-intensive tasks with similar scopes, they have not (at the time of writing) communicated or shared experiences, not to mention building on each other's work (as already noted by Winters & Netscher, 2016).

3 The proposed workflow

In the proposed workflow harmonization documentation is a product of the harmonization process itself. This section describes the harmonization steps and associated tools, to provide an overview of the whole process.

The procedure starts with all the source data files downloaded to a single location, and their origin and versions noted. Next, each source data file is imported to R, and inspected to create the following technical variables: source case ID, target case ID (row number in the data file), survey project, survey wave/round identifier, survey year, survey country, and case weights.

In terms of data processing, the harmonization work involves working with data at different levels. For each of these steps, a corresponding table is created on the basis of the source data, and exported to a spreadsheet program for manual mapping. The resulting crosswalks are used in the next step of harmonization, and at the same time serve as documentation.

Crosswalks are commonly used tools for mapping of one schema onto the other. They are most useful when the source schema can be unambiguously translated into the target schema. This is rarely the case as exemplified by the many examples of class schemas developed in the social sciences. Also in the process of ex-post survey data harmonization, in some cases the mapping is not straightforward and decisions, sometimes arbitrary, need to be made. Like in other cases, also here, “[T]he key to a successful metadata crosswalk is intelligent flexibility” (Hillmann & Westbrooks, 2004, p. 91). This is why mapping is performed manually. Given the typical number of variables harmonized, this should not be an excessive effort.

Step 1: Selection of source variables for harmonization,
Step 2: Mapping source values to target values,
Step 3: Recording characteristics of source items.

The first step is at the variable level, and involves identifying variables for harmonization in the source data files, and assigning a standardized target variable name to each source variable of interest. To do this, a table with a list of all source variables is created on the basis of metadata from the source data files, and includes variable names, labels, as well as values and value labels and their corresponding frequencies. Each variable corresponds to a single row in the resulting table, which I refer to as the *codebook*. The codebook is then exported to a spreadsheet program, where variable lables are filtered to inentify candidates for source variables. In some cases, for example is variable labels are too short or otherwise uninformative - the original survey documentation needs to be consulted. After all source variables corresponding to the concept of interest - in this case trust in institutions - are identified via filter searches or otherwise, the table is imported back into R.

The second step is at the level of individual values of the source variables, and requires mapping these source values onto a common coding scheme of the respective target variable. Subsets of source data files are selected that include only the variables tagged in Step 1. For those variables, a *cross-walk table* is created, i.e., a table similar to the codebook, but where each source value corresponds to one row. This table is then exported to a spreadsheet, where each source value is assigned a target value on the basis of the common coding scheme. This step also includes identifying missing value codes (such as negative numbers or multiples of 8 or 9) as missing. Next, the cross-walk table is imported back to R, where the source and target values for each source variable are mapped resulting in the harmonized variables.

The third step is at the data file level, and refers to the coding of properties of the source variables that are worth preserving because of methodological reasons, such as the length or direction of the original response scales. Typically, these properties vary between, but not within, survey projects. For example, in the European Social Survey, questions about trust in institutions are accompanied by 11-point scales ranging from 0 (No trust at all) to 10 (Completely trust) (European Social Survey, 2016).

4 Illustration: Trust in institutions

The illustrative example deals with trust in institutions items and their basic correlates in three main cross-national survey projects in Europe: the European Social Survey, the European Values Study, and the European Quality of Life Survey. The data are available from the project website (European Social Survey, 2018), the GESIS archive (European Values Study, 2015, 2018), and the UK Data Archive (European Quality of Life Survey, 2018), respectively. All three projects conduct surveys in many European countries in each project wave, with samples intended as representative for entire adult populations of the respective country. Table 1 presents basic information about these survey projects. Standardized documentation about the methodology employed in all projects has been put together by Piotr Jabkowski (2018).

Trust in political and other institutions is a common item in many cross-national surveys, and appears in all waves of all three projects. Each project includes questions about trust in a different set of institutions. Further, while in all three projects questionnaires tend to be relatively stable from wave to wave, the sets of questions have seen some changes, in particular new items added, and not always a given item was asked in all countries in a given wave. Finally, in each project the design of the trust items is slightly different. Major differences include the length and direction of the response scale, while other differences - response scale polarity or details of the question wording - seem to be of lesser importance for the distribution of the answers (Kołczyńska & Slomczynski, 2018).

For researchers interested in analyzing trust in a comparative perspective, it is important to know which projects include which trust items in which waves and countries, i.e., how rich a dataset they can count on. Like in all situations where the availability of variables for analysis is not clear from the start, this is an exploratory situation.

Table 1: Description of the survey projects.

	Number of			
Project name	waves	data files	surveys	Years
European Social Survey (ESS)	8	1	195	2002-2017
European Quality of Life Survey (EQLS)	4	1	126	2003-2016
European Values Study (EVS)	4	2	140	1981-2017

4.1 Step 0: Preparation and coding of technical variables

In the proposed schema, survey projects correspond to the organizational structures that public data under the same brand, such as the World Values Survey or the European Social Survey. Some of the projects have multiple waves of data collection. Each wave consists of national surveys, identified as a survey carried out in a given country and project wave. National surveys are identified as *project*wave*country*, which typically is equivalent to *project*year*country*. Most often the national survey level is equivalent to the sample level, i.e., a national survey is administered to respondents who are part of the same sample, with the sample representative for the entire population of a given country.[6]

In the preparatory step, all datafiles are loaded into the R workspace. The SPSS file format is preferred, since it comes with variable and value names that are easy to extract. Next, the so called *technical variables* are identified, corresponding to the survey wave (where applicable), country, year, case (respondent) IDs, and weighting factors. A new variable corresponding to the name of the source table is created, as well as a `wave` variable in the case of single-wave files.

[6] In some cases, in more than one sample is drawn in one country for historical, administrative, or cultural reasons, e.g., in Germany (East and West Germany) or in Belgium (Wallonia and Flanders). These sub-national samples can be either treated as separate surveys, or combined into a single survey with appropriate weights. In the projects included in this analysis the problem of sub-national samples does not exist, but it is signaled for the sake of clarity.

4.2 Step 1: Selection of source variables for harmonization

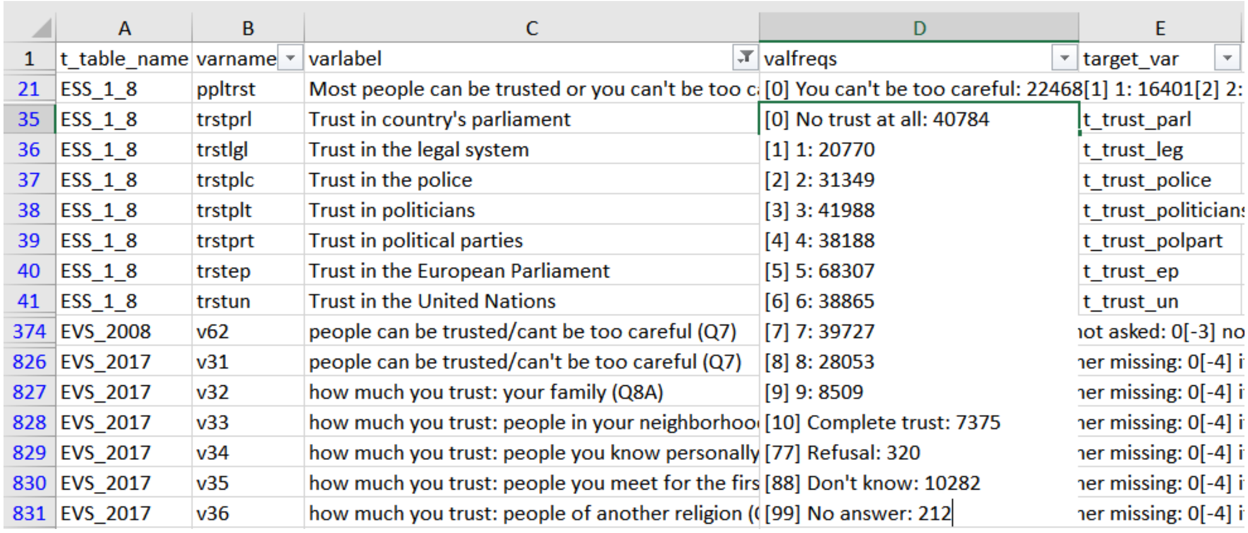
The first step in the harmonization process is the selection of source variables corresponding to the target concepts of interest. The target concepts in this case include trust in different institutions, as many as there are available in the source data.

To map source variables to target variable names, a list of source variables in each dataset is necessary, which I refer to as the *codebook*. This list should include the source variable name, and - since variable names tend to by cryptic - also the source variable label. The frequency distribution of values, as well as value labels, provide additional information in case the variable label is not informative, e.g., to distinguish age in years from age in categories.

The codebook can be generated from labelled datasets, such as those created in SPSS. In R, such a codebook can be created with the package `labelled` (Larmarange, 2019). The R code for creating the codebook is presented in Appendix 1.

The codebook is the exported into a spreadsheet program (e.g., with the `rio` package (Chan, Chan, Leeper, & Becker, 2018)), where variable labels can be browsed and filtered to select the source variables of interest, and assign appropriate target variable names. The target variable names will be `t_trust_*`, where `*` is the short version of the institution name, and the prefix `t_` identifies the target variable. If necessary, notes or comments on particular decisions can be written in a separate column(s). Notes may document decisions to treat questions about slightly different wording as corresponding to the same target variable (e.g., trust in the justice system vs. trust in the legal system), or add information from other sources, such as the original documentation of the source data.

Figure 1 shows a snippet of the codebook created for ESS, EVS, and EQLS. The first column (`t_table_name`) corresponds to the name of the source table - one each for ESS and EQLS, and two separate tables for EVS/5 and EVS/1-4. The column `varname` contains the names of source variables, `varlabel` contains the variable labels, and `valfreqs` contains frequencies of all values with value labels provided. The column `target_var` is to be filled with names of target variables to which the selected source variables correspond. In the example shown in



	A	B	C	D	E
1	t_table_name	varname	varlabel	valfreqs	target_var
21	ESS_1_8	ppltrst	Most people can be trusted or you can't be too c	[0] You can't be too careful: 22468[1] 1: 16401[2] 2:	
35	ESS_1_8	trstprl	Trust in country's parliament	[0] No trust at all: 40784	t_trust_parl
36	ESS_1_8	trstlgl	Trust in the legal system	[1] 1: 20770	t_trust_leg
37	ESS_1_8	trstplc	Trust in the police	[2] 2: 31349	t_trust_police
38	ESS_1_8	trstplt	Trust in politicians	[3] 3: 41988	t_trust_politicians
39	ESS_1_8	trstprt	Trust in political parties	[4] 4: 38188	t_trust_polpart
40	ESS_1_8	trstep	Trust in the European Parliament	[5] 5: 68307	t_trust_ep
41	ESS_1_8	trstun	Trust in the United Nations	[6] 6: 38865	t_trust_un
374	EVS_2008	v62	people can be trusted/cant be too careful (Q7)	[7] 7: 39727	ot asked: 0[-3] no
826	EVS_2017	v31	people can be trusted/can't be too careful (Q7)	[8] 8: 28053	er missing: 0[-4] i
827	EVS_2017	v32	how much you trust: your family (Q8A)	[9] 9: 8509	er missing: 0[-4] i
828	EVS_2017	v33	how much you trust: people in your neighborhoo	[10] Complete trust: 7375	er missing: 0[-4] i
829	EVS_2017	v34	how much you trust: people you know personally	[77] Refusal: 320	er missing: 0[-4] i
830	EVS_2017	v35	how much you trust: people you meet for the firs	[88] Don't know: 10282	er missing: 0[-4] i
831	EVS_2017	v36	how much you trust: people of another religion ((	[99] No answer: 212	er missing: 0[-4] i

Figure 1: Codebook table.

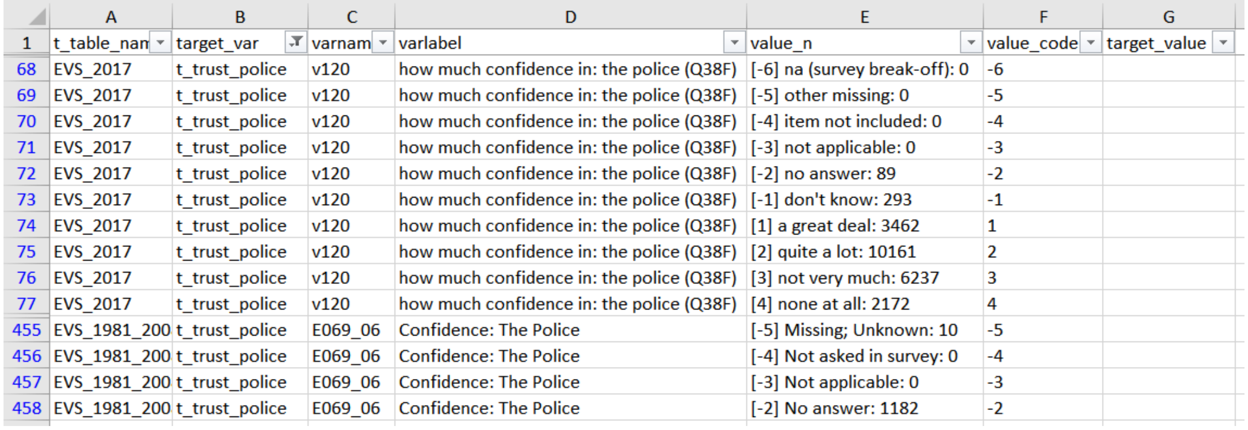
Figure 1, the label `t_trust_parl` has been assigned to the variable on trust in parliament, the label `t_trust_leg` to the variable on trust in the legal system, and so on. Filtering available in spreadsheet programs enables sorting through variables with keywords.

4.3 Step 2: Mapping source values to target values

In the second step, for the variables selected with the codebook, a *crosswalk* template is created, where each value of each variable is in a separate row. Again, the `labelled` package can be used to obtain the desired format, as shown in Appendix 2.

This manual mapping of source values to target values, while mundane, achieves two goals. First, it enables the identification of missing value codes in each variable separately. Missing value codes are not always consistent across variables even in the same dataset, and mistakes in automated recoding of missing values have a potentially large impact on the resulting data. Additionally, researchers might want to treat different missing values differently, i.e., considering the "don't know" category as a type of opinion (or rather lack thereof), and the "refusal" category as in fact missing.

The second goal is a careful inspection of full labels of the source values, one by one,



	A	B	C	D	E	F	G
1	t_table_nam	target_var	varnam	varlabel	value_n	value_code	target_value
68	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-6] na (survey break-off): 0	-6	
69	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-5] other missing: 0	-5	
70	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-4] item not included: 0	-4	
71	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-3] not applicable: 0	-3	
72	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-2] no answer: 89	-2	
73	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[-1] don't know: 293	-1	
74	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[1] a great deal: 3462	1	
75	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[2] quite a lot: 10161	2	
76	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[3] not very much: 6237	3	
77	EVS_2017	t_trust_police	v120	how much confidence in: the police (Q38F)	[4] none at all: 2172	4	
455	EVS_1981_200	t_trust_police	E069_06	Confidence: The Police	[-5] Missing; Unknown: 10	-5	
456	EVS_1981_200	t_trust_police	E069_06	Confidence: The Police	[-4] Not asked in survey: 0	-4	
457	EVS_1981_200	t_trust_police	E069_06	Confidence: The Police	[-3] Not applicable: 0	-3	
458	EVS_1981_200	t_trust_police	E069_06	Confidence: The Police	[-2] No answer: 1182	-2	

Figure 2: Crosswalk table.

while consulting the textual survey documentation where necessary. The need to assign each source value a corresponding target value again protests from some mistakes of automation, and is particularly useful in the case of categorical variables with limited standardization across waves and projects, such as education or size of town of residence.

The snippet in Figure 2 shows a fragment of the crosswalk for trust in the police in EVS/Round 5, for which the source variable is called `v120`, and realized values include negative codes for different types of missing data, and values from 1 to 4 corresponding to substantive answers. Corresponding target values would be added in the `target_value` column.

Once the value crosswalk table is ready, it is imported into R, and - for each source variable separately - source and target values are extracted as vectors and used in mapping. The function `mapvalues` from the `plyr` package (Wickham, 2011), which maps each value from the source variable vector onto a corresponding value from the target variable vectors, is particularly useful in this regard.

4.4 Step 3: Recording characteristics of source items

Cross-national surveys differ in almost all impaginable aspects related to the methodology of the survey process, starting from questionnaire development and pretesting, through fieldwork procedures and response or realization rates, the design of individual survey items (Kołczyńska &

Slomczynski, 2018), to the quality of documentation (Jabkowski & Kołczyńska, 2019; Kołczyńska & Schoene, 2018). According to the literature on survey methodology, in particular in the Total Survey Error paradign, the comparability of surveys is reduced whenever the amounts of random and systematic error are different across surveys (Smith, 2018). The TSE framework currently counts 35 different sources of error, so the potential for deviations is large.

This is why, while the amount and type (random or systematic) of each type of error in surveys is unknown (and for the most part unknowlable), one approach to dealing with the methodological variation of surveys comprising a single harmonized dataset is to construct additional variables that capture selected properties of the source data or documentation in form of control variables. It is not clear whether including these control variables in substantive analyses improves model estimates or not, and under which cirsumstances, but such methodological controls can at least be used in data exploration.

In the case of questions about trust in institutions, the most easily identifiable differences across items from different projects are in the response scales, in particular in their length and direction (cf. Kołczyńska & Slomczynski, 2018). All three projects are consistent with their item design, i.e., all trust in institutions items have the same response scales within projects: the European Social Survey has 0-10 ascending scales with labels only for extreme values, the European Quality of Life Survey has 1-10 ascending scales also with only extreme points labelled, and the European Values Study has 4-point descending fully-labelled scales. In this case item designs overlap perfectly with projects, which makes the two effects - the effects of being part of aparticular organizational structure and the effects of the response scale - cannot be analytically distinguished.

Item and other methodological properties can be recorded in a separate table, where the national survey (project*wave*country) is the unit of observation.

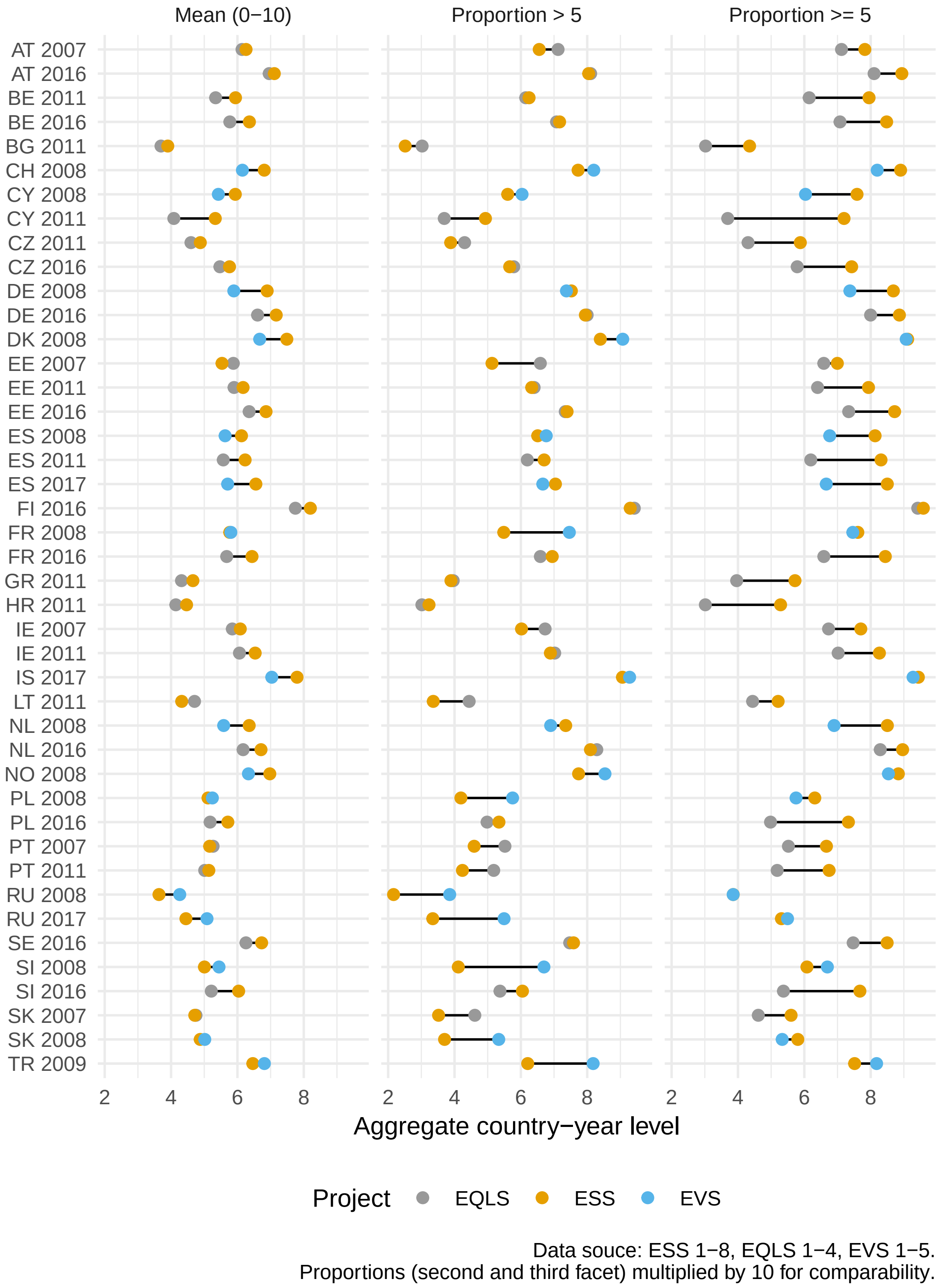


Figure 3: Country-year levels of trust in the police by aggregate type.

4.5 Results: Comparing average levels of trust in the police

5 Limitations

The presented approach naturally has some limitations. Some of them are inherent to the main tool used - the crosswalk. If the source and target schemas do not map onto each other very well, the need for the aforementioned “intelligent flexibility” (Hillmann & Westbrooks, 2004, p. 91) arises. At the same time, crosswalks enable mapping from a single scheme onto another single scheme. Some researchers might want to construct the target variable on the basis of two or more source variables. For example, the target variable “membership in organizations” could be constructed on the basis of two target variables corresponding to membership in political organizations and membership in non-political organizations. Cases when two (or more) variables need to be combined into one target variable can be handled easily in the code, with the identification of missing values and perhaps preliminary recoding in the crosswalk.

Another limitation deals with ex-post survey data harmonization in general. Each time a transformation of the source variable(s) is performed, the costs and benefits need to be considered. Since at least the 1940s it has been known that the design of the survey items influences the distribution of respodents’ answers, and - consequently - of sample aggregates, such as means or proportions (Cantril, 1944). In the case of items that are designed as a scale to measure a latent trait, selecting some items of the scale, or changing their coding, may undermine the validity of the scale (Mustillo, Lizardo, & McVeigh, 2018).

While the literature on question and questionnaire design effects is rich and constantly growing, recommendations are typically formulated with regard to best practices in future data collection efforts, not from the point of view of procedures that improve the comparability of already collected data.

6 Recommendations

An absolute minimum for reproducibility of any secondary analysis or processing of survey data is the information about:
1. The origin of the data files, their format, version, and information about how to obtain them, 2. Original data files (unless not possible due to legal or ethical restrictions), 3. Original data documentation, with their version and origin recorded, 4. Recoding syntax in a recognized language, 5. Instructions describing the procedures.

It is possible that this "minimum package" will actually enable the reproduction of all procedures, but it is likely that it will not. The potential reasons are many, ranging from insufficient competence in the given programming language, to errors in the code, ambiguous instructions, lack of documentation, differences in the software or hardware environment. Some problems with reproducing analyses in a relatively favorable situation with good will and communication from both sides are described by Liu and Salganik (2019). Creating reproducible documentation is hard, because missing even one of the necessary steps or components leads to the failure of the whole process.

There are several ways of improving data processing reproducibility. The first one is automating all automatable data processing steps, including importing the data, cleaning, processing, and exporting. The code should be annotated. Following Liu & Salganik (2019), code documents should include information about the general purpose of the code segment, as well as a description of the input, the outputs, the type of machine used, and the expected runtime.

All steps of manual processing should be documented, ideally in machine-readable form, with document versioning and visible corrections. The latter point is important to recostruct the decision changes, and in tracking down errors.

Chances for successful data processing reproduciblity increase if the software environement is documented, i.e. versions of all software and packages or libraries that were used in data processing.

Reproduciblity of data processing work in the long term benefits from the human-friendliness and universal format of the documentation files. In this case simpler is better, and chances are that text files will be usable "forever", which is not necessary true for SPSS syntax files. Hence, data processing syntax in other forms than just code - such as crosswalks - has a potential to be accessible more widely and for longer.

In the recent years there has been increasing awareness of the value and challenges of repricability and reproduciblity, both among researchers and educators. Reith, Paxton, and Hughes (2016) offer recommendations for creating cross-national, logitudinal datasets, many of which also apply to harmonizing survey data. Project TIER (Teaching Integrity in Empirical Research) developed the DRESS protocol - a set of guidelines for documenting empirical research (Project TIER, 2016). Liu and Salganik (2019) describe their experiences with verifying computational reproducibility of papers stemming from the Fragile Families Challenge.

7 Appendices

Functions to construct the codebook and values crosswalk, as well as the surveys list, are in Appendix 1, 2, and 3, respectively. The functions rely on the R language (R Core Team, 2018), in particular the packages `haven` for importing the data into R (Wickham & Miller, 2019), `labelled` for dealing with labels (Larmarange, 2019), and the `tidyverse` suite for data wrangling (Wickham, 2017).

7.1 Appendix 1: Codebook from labelled data in R

```
create_codebook <- function(data) {
  var_labels <- data.frame(cbind(names(var_label(data)), do.call(rbind, var_label(data)))) %>%
    rename(varname = X1, varlabel = X2)
  freqs <- lapply(data, function(x) { return(questionr::freq(x)) }) %>%
    keep(function(x) nrow(x) < 1000) %>%
    do.call(rbind, .) %>%
    tibble::rownames_to_column(var = "varname_value") %>%
    mutate(varname = gsub("(.+?)(\\..*)", "\\1", varname_value),
           value = gsub("^[^.]*.","",varname_value)) %>%
    group_by(varname) %>%
    mutate(npos = row_number(),
           value_n = paste(value, n, sep = ": ")) %>%
    select(varname, value_n, npos) %>%
    spread(npos, value_n) %>%
    mutate_at(vars(-varname), funs(ifelse(is.na(.), "", .))) %>%
    unite("valfreqs", c(2:ncol(.)), sep = "\n") %>%
    mutate(valfreqs = sub("\\s+$", "", valfreqs))
  full_join(var_labels, freqs, by = "varname")
}
```

7.2 Appendix 2: Values crosswalk

```
create_cwt <- function(data) {
  var_labels <- data.frame(cbind(names(var_label(data)), do.call(rbind, var_label(data)))) %>%
    rename(varname = X1, varlabel = X2)
  freqs_cwt <- lapply(data, function(x) { return(questionr::freq(x)) }) %>%
    keep(function(x) nrow(x) < 1000) %>%
    do.call(rbind, .) %>%
    tibble::rownames_to_column(var = "varname_value") %>%
    mutate(varname = gsub("(.+?)(\\..*)", "\\1", varname_value),
           value = gsub("^[^.]*.","",varname_value),
           value_code = sub(".*\\[(.+)\\].*", "\\1", varname_value, perl = TRUE),
           value_code = ifelse(str_sub(varname_value, -2, -1) == "NA", "NA", value_code),
           value_code = ifelse(gsub(" ", "", fixed = TRUE, varname_value) == varname_value,
                               gsub("^[^.]*.","",varname_value), value_code)) %>%
    group_by(varname) %>%
    mutate(npos = row_number(),
           value_n = paste(value, n, sep = ": ")) %>%
    select(npos, varname, value_n, value, value_code)
  full_join(var_labels, freqs_cwt, by = "varname")
}
```

7.3 Appendix 3: Metadata table

```
# This function can be run after the following variables in the source data have been constructed:
# t_project for the project abbreviation;
# t_round for the project round;
# t_country for the country code;
# t_year for the survey year.

create_surveys_list <- function(data_frame_vector) {
  surveys_list <- list()
  for (i in 1:length(data_frame_vector)) {
    surveys_list[[i]] <- eval(parse(text = data_frame_vector[i])) %>%
      mutate(t_survey = paste(t_project, t_round, t_country, sep = "")) %>%
      count(t_survey, t_year)
  }
  do.call(rbind, surveys_list)
}
```

8 References

Bekkers, R., Meer, T. van der, Uslaner, E., Wu, Z., Wit, A. de, & Blok, L. de. (2015). Harmonized Trust Database. https://doi.org/10.17605/OSF.IO/QFV76

Buckheit, J. B., & Donoho, D. L. (1995). WaveLab and Reproducible Research. In A. Antoniadis & G. Oppenheim (Eds.), *Wavelets and statistics* (pp. 55–81). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4612-2544-7_5

Burkhauser, R. V., & Lillard, D. R. (2005). The contribution and potential of data harmonization for cross-national comparative research. *Journal of Comparative Policy Analysis: Research and Practice*, *7*(4), 313–330. https://doi.org/10.1080/13876980500319436

Cantril, H. (1944). *Gauging Public Opinion*. Princeton, NJ: Princeton University Press.

Chan, C.-h., Chan, G. C. H., Leeper, T. J., & Becker, J. (2018). *rio: A Swiss-army knife for data file I/O*.

Christmann, P. (2018). Economic performance, quality of democracy and satisfaction with democracy. *Electoral Studies*, *53*(April 2017), 79–89. https://doi.org/10.1016/j.electstud.2018.04.004

Christmann, P., & Torcal, M. (2018). The Effects of Government System Fractionalization on Satisfaction with Democracy. *Political Science Research and Methods*, *6*(3), 593–611. https://doi.org/10.1017/psrm.2017.23

Curran, P. J., & Hussong, A. M. (2009). Integrative Data Analysis: The Simultaneous Analysis of Multiple Data Sets. *Psychological Methods*, *14*(2), 81–100. https://doi.org/10.1037/a0015914

European Quality of Life Survey. (2018). European Quality of Life Survey Integrated Data File, 2003-2016. European Foundation for the Improvement of Living; Working Conditions. UK Data Service. https://doi.org/10.5255/UKDA-SN-7348-3

European Social Survey. (2016). ESS Round 8 Source Questionnaire. London: ESS ERIC Headquarters c/o City University London.

European Social Survey. (2018). European Social Survey Cumulative File, ESS 1-8 (2018). NSD - Norwegian Centre for Research Data, Norway - Data Archive; distributor of ESS data for ESS ERIC. https://doi.org/10.21338/NSD-ESS-CUMULATIVE

European Values Study. (2015). European Values Study Longitudinal Data File 1981-2008 (EVS 1981-2008). Cologne: GESIS Data Archive. https://doi.org/10.4232/1.12253

European Values Study. (2018). European Values Study 2017: Integrated Dataset (EVS 2017). Cologne: GESIS Data Archive. https://doi.org/10.4232/1.13090

Fortier, I., Raina, P., Van den Heuvel, E. R., Griffith, L. E., Craig, C., Saliba, M., … Burton, P. (2017). Maelstrom Research guidelines for rigorous retrospective data harmonization. *International Journal of Epidemiology*, *46*(1), 103–115. https://doi.org/10.1093/ije/dyw075

Granda, P., & Blasczyk, E. (2016). Data Harmonization. In *Guidelines for best practice in cross-cultural surveys* (pp. 617–635). Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. Retrieved from http://www.ccsg.isr.umich.edu/

Hillmann, D. I., & Westbrooks, E. L. (2004). *Metadata in Practice* (p. 285). American Library Association.

Jabkowski, P. (2018). *Surveys Quality Assessment Database (SQAD)*. Retrieved from https://www.researchgate.net/project/Surveys-Quality-Assessment-Database-SQAD

Jabkowski, P., & Kołczyńska, M. (2019). Survey practices in Europe: Analysis of methodological documentation from 1537 surveys in five cross-national projects, 1981-2017. Unpublished draft.

Klassen, A. J. (2018). *Human Understanding Measured Across National (HUMAN) Surveys. Codebook for Respondent Data* (No. 15 February). Harvard Dataverse, V1. https://doi.org/10.7910/DVN/QLKR85

Kołczyńska, M. (2019). From Cross-national ex post Survey Harmonization to Substantive Analyses: A Roadmap and Empirical Illustration of Micro- and Macro-level Determinants of Protest Participation. Unpublished draft.

Kołczyńska, M., & Schoene, M. (2018). Survey Data Harmonization and the Quality of Data Documentation in Cross-national Surveys. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 963–984). Wiley.

Kołczyńska, M., & Slomczynski, K. M. (2018). Item Metadata as Controls for Ex Post Harmonization of International Survey Projects. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 1011–1034). Wiley.

Larmarange, J. (2019). *labelled: Manipulating Labelled Data*. Retrieved from https://cran.r-project.org/package=labelled

Liu, D. M., & Salganik, M. J. (2019). *Successes and struggles with computational reproducibility: Lessons from the Fragile Families Challenge* (pp. 1–56). https://doi.org/10.31235/osf.io/g3pdb

Mauk, M. (2019). Disentangling an Elusive Relationship: How Democratic Value Orientations Affect Political Trust in Different Regimes. *Political Research Quarterly*. https://doi.org/10.1177/1065912919829832

Mustillo, S. A., Lizardo, O. A., & McVeigh, R. M. (2018). Editors' Comment: A Few Guidelines for Quantitative Submissions. *American Sociological Review*, *83*(6), 1281–1283. https://doi.org/10.1177/0003122418806282

Powałko, P., & Kołczyńska, M. (2016). Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions. *International Journal of Sociology*, *46*(1), 73–80. https://doi.org/10.1080/00207659.2016.1130433

Project TIER. (2016). *The DRESS Protocol: Documenting Research in the Empirical Social Sciences*. Teaching Integrity in Empirical Research (TIER). Retrieved from https://www.projecttier.org/tier-protocol/dress-protocol/

R Core Team. (2018). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Reith, N. E., Paxton, P., & Hughes, M. M. (2016). Building Cross-National, Longitudinal Data Sets: Issues and Strategies for Implementation. *International Journal of Sociology*, *46*(1), 21–41. https://doi.org/10.1080/00207659.2016.1130416

Sandberg, B., & Bekkers, R. (2018). *Harmonized Trust Database Codebook, Version 1.3*. Global Trust Research Consortium. Retrieved from https://osf.io/92r5z/

Slomczynski, K. M., Jenkins, J. C., Tomescu-Dubrow, I., Kołczyńska, M., Wysmułek, I., Oleksiyenko, O., … Zieliński, M. W. (2017). SDR Master Box. Retrieved from https://doi.org/10.7910/DVN/VWGF5Q

Slomczynski, K. M., & Tomescu-Dubrow, I. (2018). Basic Principles of Survey Data Recycling. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in comparative survey methods: Multinational, multiregional, and multicultural contexts (3MC)* (pp. 937–962). Wiley. https://doi.org/10.1002/9781118884997.ch43

Smith, T. W. (2018). Improving Multinational, Multiregional, and Multicultural (3MC) Comparability Using the Total Survey Error (TSE) Paradigm. In *Advances in comparative survey methods* (pp. 13–43). Hoboken, NJ, USA: John Wiley & Sons, Inc. https://doi.org/10.1002/9781118884997.ch2

Słomczyński, K. M., Tomescu-Dubrow, I., & Jenkins, J. C. (2016). *Democratic Values and Protest Behavior. Harmonization of Data from International Survey Projects.* Warsaw: IFiS Publishers.

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29. Retrieved from http://www.jstatsoft.org/v40/i01/

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. Retrieved from https://cran.r-project.org/package=tidyverse

Wickham, H., & Miller, E. (2019). *haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files*. Retrieved from https://cran.r-project.org/package=haven

Winters, K., & Netscher, S. (2016). Proposed Standards for Variable Harmonization Documentation and Referencing: A Case Study Using QuickCharmStats 1.1. *PLoS ONE*, *11*(2), 1–15. https://doi.org/10.1371/journal.pone.0147795

Wolf, C., Schneider, S. L., Behr, D., & Joye, D. (2016). Harmonizing Survey Questions Between Cultures and Over Time. In *The sage handbook of survey methodology* (pp. 502–524). SAGE. https://doi.org/10.4135/9781473957893