WAPOR WEBINAR SERIES, May 19 2022

Survey Data Recycling and Data Harmonization

Irina Tomescu-Dubrow, IFiS PAN and CONSIRT

Ilona Wysmulek, IFiS PAN

CONSIRT.osu.edu

Survey Data Harmonization

= body of methods to achieve/strengthen comparability of survey data

Data producers → Ex-ante/prospective harmonization

- before fieldwork (input harmonization)
- during data processing, before data release (output harmonization)

Secondary users → Ex-post/retrospective harmonization

Secondary users → ex-post/retrospective harmonization:

Methods to select variables capturing the same concept in surveys not designed as comparative, transform them to achieve/increase the comparability of respondents' answers, and create an integrated dataset that can be analyzed as a single data source.

(e.g. Günther 2003; Minkel 2004; Granda, Wolf & Hadorn 2010)

Purpose: go from **good** to **better** data infrastructure for comparative analyses without fielding new surveys

Theory & pragmatism inform the scope of ex-post harmonization

SDR Project

Substantive areas of comparative research in the SDR project:

Democracy and Political Participation

Social Capital and Political Participation

Social Capital and Wellbeing

→ cross-national survey data with ex-ante harmonized measures of peoples’ *political attitudes & behaviors, social capital,* and *wellbeing* + correlates

→ between-country and over-time variation in democracy and economic development

SDR Project

Individual survey projects with *political attitudes & behaviors, social capital,* and/or *wellbeing* measures

- Asian Barometer
- Afrobarometer
- Americas Barometer
- Arab Barometer
- Asia Europe Survey
- Caucasus Barometer
- Consolidation of Democracy
- Comparative National Elections Project
- Eurobarometer
- European Quality of Life Survey
- European Social Survey
- European Values Study
- International Social Justice Project
- International Social Survey Programme
- Latinobarometro
- Life in Transition Survey
- New Baltic Barometer
- Political Action II
- Political Action - An Eight Nation Study
- Political Participation and Equality
- Values and Political Change
- World Values Survey
- *New Europe Barometer*

SDR Project: From Good to **Better**

Integrated SDR Database v.2.0

Time span	1966-2017
Nr. countries	156
Nr. respondents	4,402,489
Nr. survey projects	23
Nr. project waves	174
Nr. national surveys	3,329
Nr. source data files	214

Pragmatic steps:

→ Criteria for selected survey projects:

- Non-commercial
- Cross-national, preferably, multi-wave
- National samples intended as representative of the adult population
- English language documentation (study description, codebook, questionnaire)
- Freely available for academic use

→ 09.2017 as end-date for data & documentation downloads in SDR 2.0.

Main **Challenges** in Ex-post Survey Data Harmonization

(I) Methodological biases and errors stemming from:

- (a) deviations from standards of documenting and preparing survey data suggested in the specialized literature (e.g. Biemer and Lyberg 2003)
- (b) inter-survey differences in properties of items measuring the same concept
- (c) harmonization procedures

(II) Information loss

(III) Ensure transparency

SDR Project: **Solutions**

SDR analytic framework develops ex-post harmonization methodology that incorporates

a) Classic definition

Methods to select variables capturing the same concept in surveys not designed as comparative, transform them to achieve/increase the comparability of respondents’ answers, and create an integrated dataset that can be analyzed as a single data source.

+

b) Methods to define and measure inter-survey methodological differences stemming from: (i) source survey quality
(ii) ex-post harmonization

SDR: Survey Data **Recycling**

In the SDR database:

- **Harmonized technical variables**

  (e.g. project name, interview year, survey year, country, weights)
- **Harmonized substantive variables**
- **Methodological controls for harmonization**
- **Methodological controls for source survey quality**

→ Recycling,
along 2 lines:

a) reusing data from extant international projects

b) increasing research flexibility via methodological indicators

Harmonization Workflow

Main steps of SDR *ex-post* harmonization workflow:

Step one: THEORY

Outline theoretical model(s) and concepts & define the criteria for data selection

Step two: SOURCE CHECK

Check source variables availability & make first (inclusive) selection of sources

Step three: DEFINE TARGET

(Re)define target variable based on source variables --> final source variable list

Step four: CONTROL VARIABLES

Create harmonization control variables & quality control variables

Step five: CROSS-WALK CODING

Recode values of source variables into a target → harmonization syntax

Step six: FINAL DATASET & REPORTS

Finalize integrated dataset & documentation;
detailed & general variable reports;
SDR portal and dissemination

SDR Harmonization Workflow

Step one: THEORY

Aim:

- Decide on **theoretical model(s), concepts, hypotheses** & a (desirable) list of target variables
- Conduct research on **different measures** & operationalizations of the same concept
- Define the **scope of data sources** [where & when] and concepts [narrow vs. extended] to be harmonized within those data

SDR Harmonization Workflow

Step one: THEORY

What does it mean in practice?

→ Create a (desired) list of variables for harmonization (= target variables)

→ Remain critical and open-minded to different operationalizations of the same (theoretical) target concept

→ Decide on geographical and temporary limits

→ Think of the criteria for survey project selection

→ Decide on „Time Zero” for data and documentation download

SDR Harmonization Workflow

Step two: SOURCE CHECK

Aim:

Systematic review of questionnaires and codebooks of international public opinion surveys in search for source variables that match the [theoretical] target concept.

Practical implications:

- It is helpful to define & discuss key words at this stage
- Remember to look through all available documentation sources to gain the best possible understanding of the source variable
- Decide on source quality checks that you want to make prior to harmonization

SDR Harmonization Workflow

Step two: SOURCE CHECK

SDR Tools:

→ Cotton file (Excel)= cumulative automatically-retrieved **file with list of all variables** in surveys of the SDR Database for search across original (source) datafiles

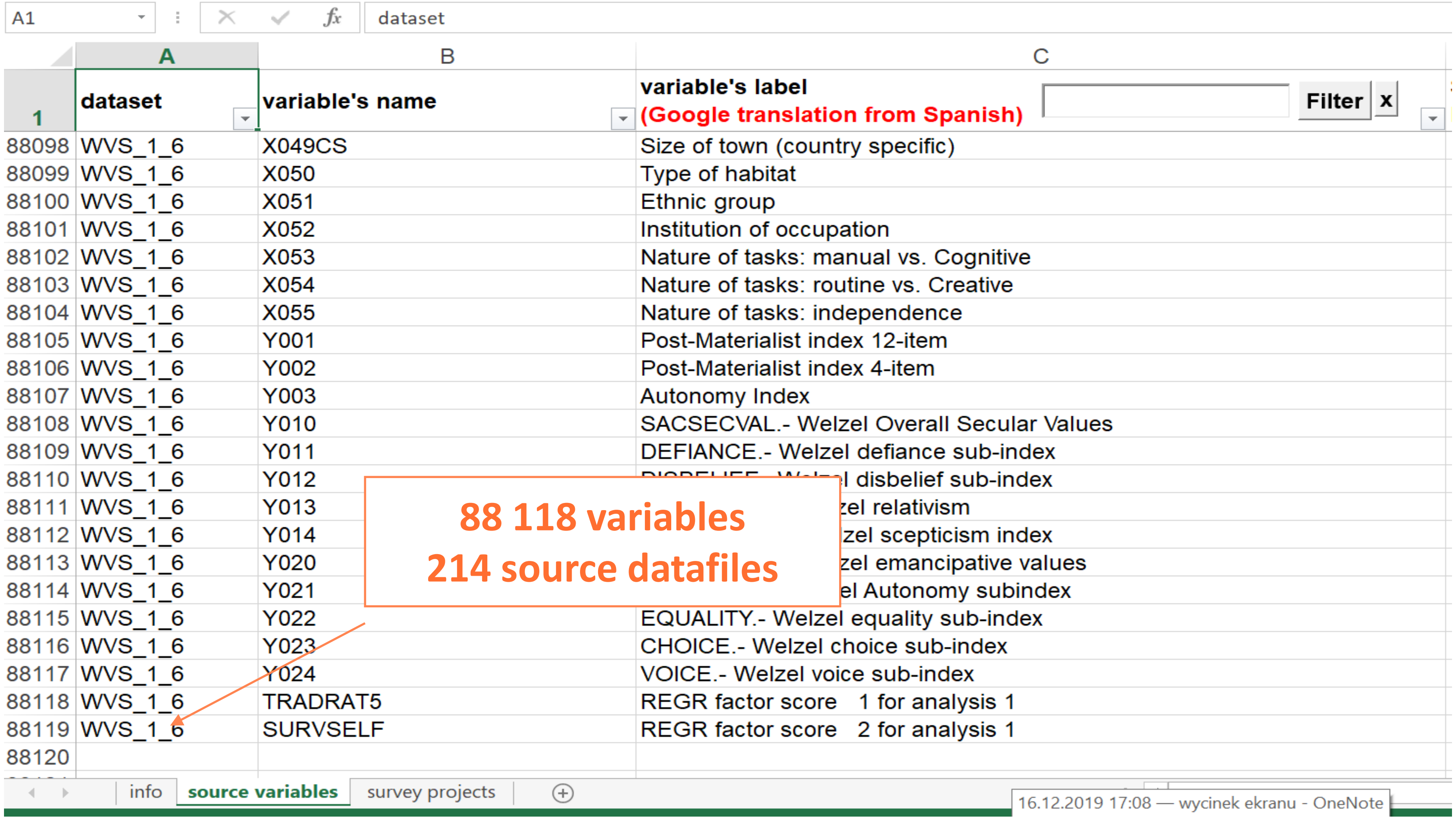
SDR 2.0 Cotton Excel file has 88,118 source variables (names/values/labels)

→ DVR (**Detailed Variable Report**, Excel) = **selected list of source variables** with harmonization potential with the standardized documentation of source variables across datasets

SDR 2.0 has DVR files corresponding to all harmonized variables

SDR Harmonization Workflow

SDR 2.0 COTTON FILE



1	dataset	variable's name	variable's label (Google translation from Spanish)
88098	WVS_1_6	X049CS	Size of town (country specific)
88099	WVS_1_6	X050	Type of habitat
88100	WVS_1_6	X051	Ethnic group
88101	WVS_1_6	X052	Institution of occupation
88102	WVS_1_6	X053	Nature of tasks: manual vs. Cognitive
88103	WVS_1_6	X054	Nature of tasks: routine vs. Creative
88104	WVS_1_6	X055	Nature of tasks: independence
88105	WVS_1_6	Y001	Post-Materialist index 12-item
88106	WVS_1_6	Y002	Post-Materialist index 4-item
88107	WVS_1_6	Y003	Autonomy Index
88108	WVS_1_6	Y010	SACSECVAL.- Welzel Overall Secular Values
88109	WVS_1_6	Y011	DEFIANCE.- Welzel defiance sub-index
88110	WVS_1_6	Y012	DISBELIEF.- Welzel disbelief sub-index
88111	WVS_1_6	Y013	zel relativism
88112	WVS_1_6	Y014	zel scepticism index
88113	WVS_1_6	Y020	zel emancipative values
88114	WVS_1_6	Y021	el Autonomy subindex
88115	WVS_1_6	Y022	EQUALITY.- Welzel equality sub-index
88116	WVS_1_6	Y023	CHOICE.- Welzel choice sub-index
88117	WVS_1_6	Y024	VOICE.- Welzel voice sub-index
88118	WVS_1_6	TRADRAT5	REGR factor score 1 for analysis 1
88119	WVS_1_6	SURVSELF	REGR factor score 2 for analysis 1
88120			

88 118 variables
214 source datafiles

SDR Harmonization Workflow

Step three: DEFINE TARGET

Aim:

- discuss (and test) strength and weaknesses of available operationalizations of a target concept and select source variables that will be matched to target
- (re)define target variable based on available source variables

What does it mean in practice?

→ Data availability may force the redefinition of a target concept

→ Team discussions are helpful – prepare a handout summing up availability of source variable types

→ Discuss which source variables stretch the target concept too much

→ Move these auxiliary variables to the „Left over" file (e.g. Excel sheet) – leave traces of your decisions

SDR Harmonization Workflow

SDR database v.2.0: Target variables

26 harmonized (target) variables in SDR 2.0

Types of variables:

Technical variables – such as project name, interview year, survey year, country, weights)

Substantive variables measuring respondents' characteristics

(a) socio-demographics - e.g. age, gender, martial status, education, income, place of residence
(b) reported behaviors - e.g. participation in demonstrations, membership in organizations)
(c) attitudes and opinions - e.g. institutional trust, trust in people, life satisfaction

+ *Missing codes schema* – standardized across all harmonized variables
+ *Harmonization control variables*

SDR Harmonization Workflow

Step four: CONTROL VARIABLES

Methodological **Controls: Harmonization** Process -- target variable specific

Preserve features of (a) the source items and (b) harmonization, which could introduce methodological differentiation:

Ex: (a) differences in meaning of the source variables (scope, time, space, etc.)
diffr. in formal properties of scale measures (scale length, direction, polarity)
diffr. number of source variables used to construct a target variable

(b) some values of a target variable are derived (e.g. age derived from birth year)

Prevent information loss and aid transparency

SDR Harmonization Workflow

Step four: CONTROL VARIABLES

Methodological **Controls**: **Source Data Quality**

Capture variability in source data quality, defined in terms of:

(i) survey documentation (codebook, questionnaires, technical reports, etc.)

inadequate information in documentation reduces the data’s fitness for use

(ii) data records in computer files

measurement & representation errors can lead to distortion of empirical results

(iii) consistency documentation – data records (for subset of variables)

processing errors can affect the overall usability of the survey

SDR Harmonization Workflow

Step five: CROSS-WALK CODING

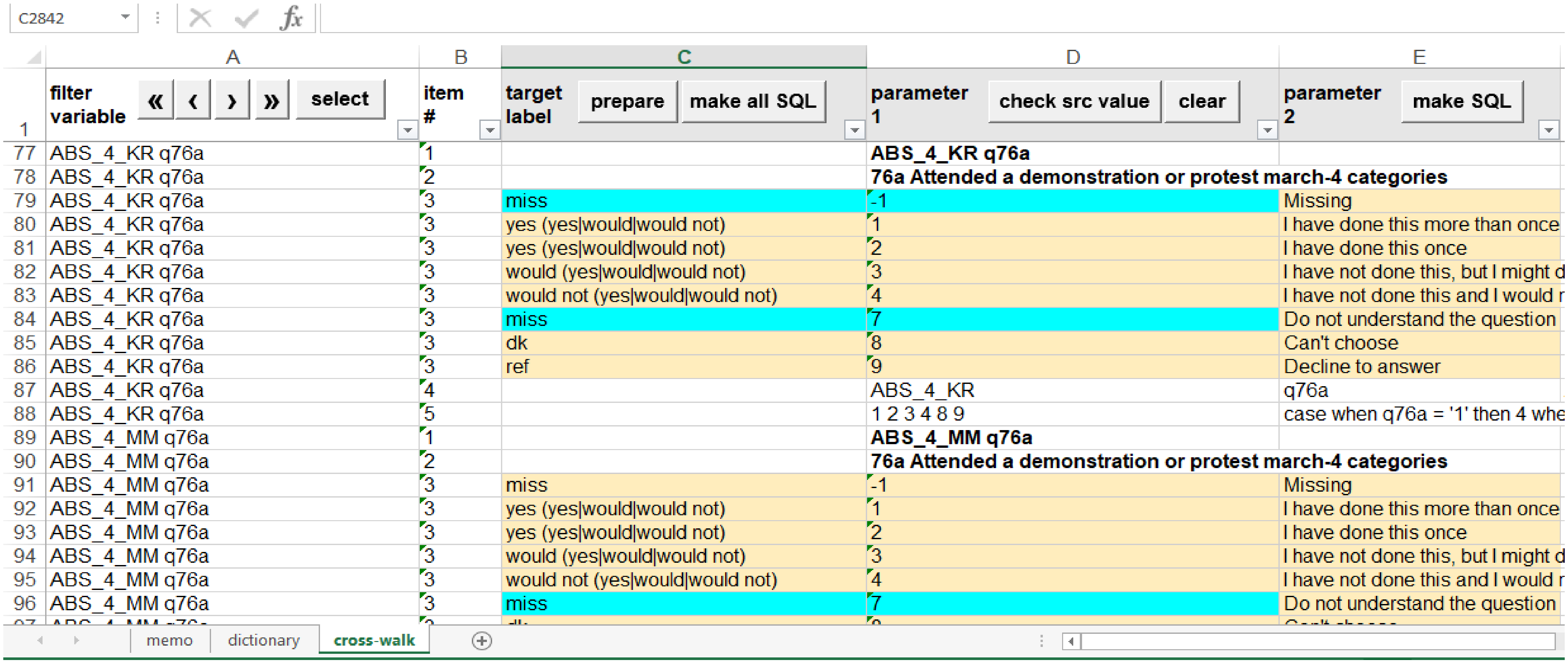
AIM: Maping source variables to target variables

CWTs or Cross-Walk Tables are macro-enabled Excel documents, the basis for harmonization syntax

→ Contain detailed information on recodes from source to target values;
→ Serve for additional consistency checks and enables quick insight into data (national survey level distributions for each source variable).

SDR Harmonization Workflow

SDR 2.0 CROSS-WALK TOOL



	A	B	C	D	E
1	filter variable « ‹ › » select	item #	target label prepare make all SQL	parameter 1 check src value clear	parameter 2 make SQL
77	ABS_4_KR q76a	1		**ABS_4_KR q76a**	
78	ABS_4_KR q76a	2		**76a Attended a demonstration or protest march-4 categories**	
79	ABS_4_KR q76a	3	miss	-1	Missing
80	ABS_4_KR q76a	3	yes (yes|would|would not)	1	I have done this more than once
81	ABS_4_KR q76a	3	yes (yes|would|would not)	2	I have done this once
82	ABS_4_KR q76a	3	would (yes|would|would not)	3	I have not done this, but I might d
83	ABS_4_KR q76a	3	would not (yes|would|would not)	4	I have not done this and I would r
84	ABS_4_KR q76a	3	miss	7	Do not understand the question
85	ABS_4_KR q76a	3	dk	8	Can't choose
86	ABS_4_KR q76a	3	ref	9	Decline to answer
87	ABS_4_KR q76a	4		ABS_4_KR	q76a
88	ABS_4_KR q76a	5		1 2 3 4 8 9	case when q76a = '1' then 4 whe
89	ABS_4_MM q76a	1		**ABS_4_MM q76a**	
90	ABS_4_MM q76a	2		**76a Attended a demonstration or protest march-4 categories**	
91	ABS_4_MM q76a	3	miss	-1	Missing
92	ABS_4_MM q76a	3	yes (yes|would|would not)	1	I have done this more than once
93	ABS_4_MM q76a	3	yes (yes|would|would not)	2	I have done this once
94	ABS_4_MM q76a	3	would (yes|would|would not)	3	I have not done this, but I might d
95	ABS_4_MM q76a	3	would not (yes|would|would not)	4	I have not done this and I would r
96	ABS_4_MM q76a	3	miss	7	Do not understand the question
97	ABS_4_MM q76a	3	dk	8	Can't choose

SDR Harmonization Workflow

Step six: FINAL DATASET & REPORTS

General target variable report (Word doc)

Detailed variable report (Excel)

Crosswalk table (macro-enabled Excel)

Syntax (notepad++)

for each Target variable & Harmonization controls

Documentation (Word doc)

Syntax (notepad++)

Source data quality controls

SDR 2.0 Cotton file

Overview of 88,118 variable names, values, and labels available in the original (source) data files that we retrieved automatically for harmonization purposes in the SDR Project

(asc.ohio-state.edu/dataharmonization/data/sdr-2-0-cotton-file)

SDR Project: Harmonization Workflow

Tools & Traces:

Step one: THEORY

-> Time zero: data and documentation downloaded

Step two: SOURCE CHECK

COTTON FILE
(Excel, source data dictionary)

Step three: DEFINE TARGET

DETAILED VARIABLE REPORT
(Excel, harmonization DVRx)

Step four: CONTROL VARIABLES

- DETAILED VARIABLE REPORT &

- QUALITY CONTROLS PLUG(s)
(Excel & data file]

Step five: CROSS-WALK CODING

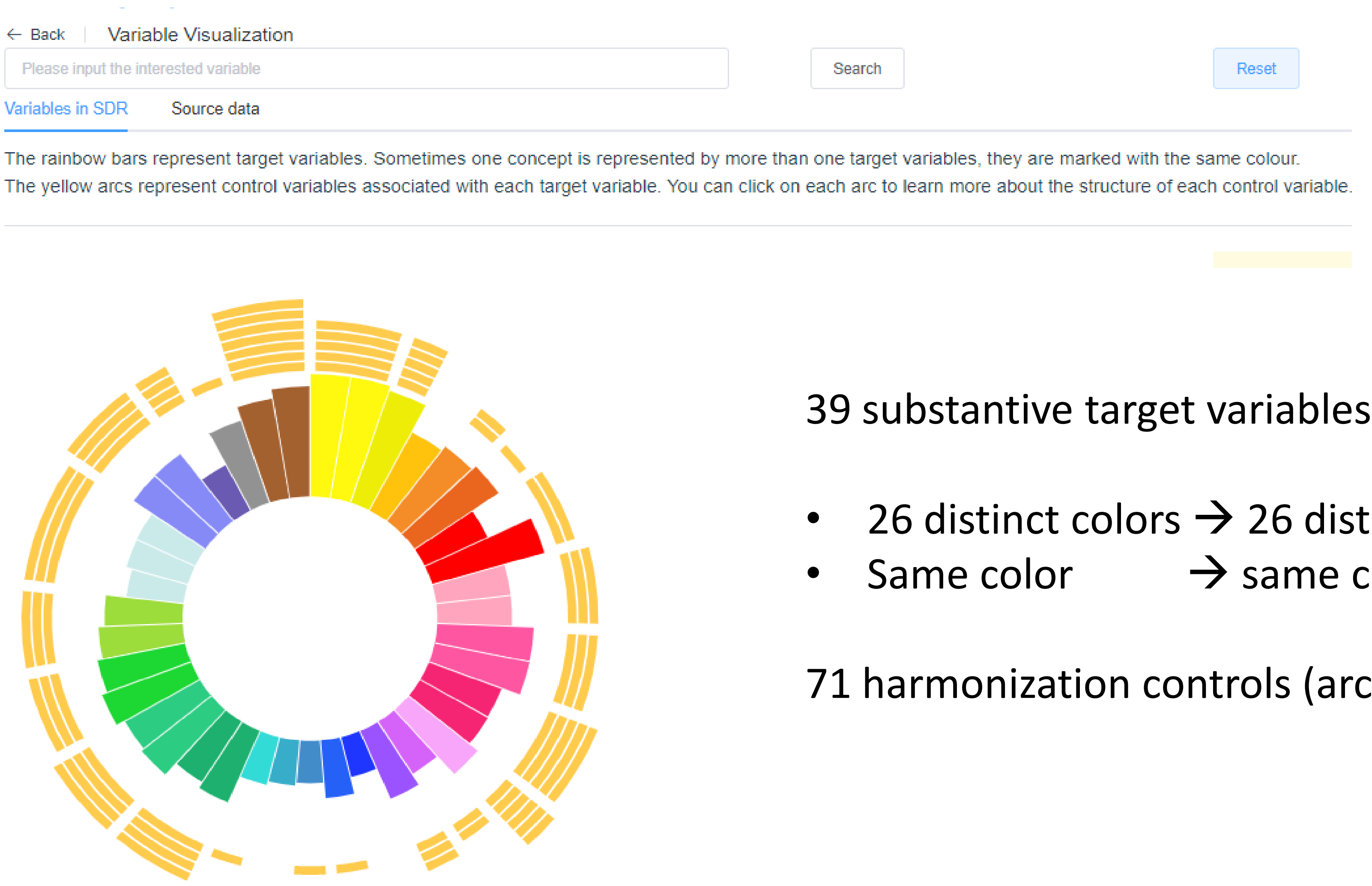
CROSS-WALK TABLE

(Excel, harmonization CWT)

Step six: FINAL DATASET & REPORTS

- MASTERFILE (Harmonized dataset)
- DATA USER GUIDE (i.a. General variable reports)

SDR Portal: illustration



The rainbow bars represent target variables. Sometimes one concept is represented by more than one target variables, they are marked with the same colour.
The yellow arcs represent control variables associated with each target variable. You can click on each arc to learn more about the structure of each control variable.

39 substantive target variables (T), by color:

- 26 distinct colors → 26 distinct concepts
- Same color → same concept, different T

71 harmonization controls (arcs)

Yamei Tu, OSU; Przemek Powalko, IFiS; Olga Li, GSSR IFiS; Zuzanna Skora, IFiS

Concluding remarks

Ex-post survey data harmonization can really be a useful tool if theory-informed

Use theory + pragmatism to set the boundaries of the ex-post harmonization project

Maximize research potential of the harmonized dataset:

- inclusive structure of target & harmonization control variables
- store source survey and harmonization metadata as methodological variables

Transparency of harmonization workflow facilitates knowledge cumulation

Needed: methodology to define & assess quality of harmonized datasets

Acknowledgements

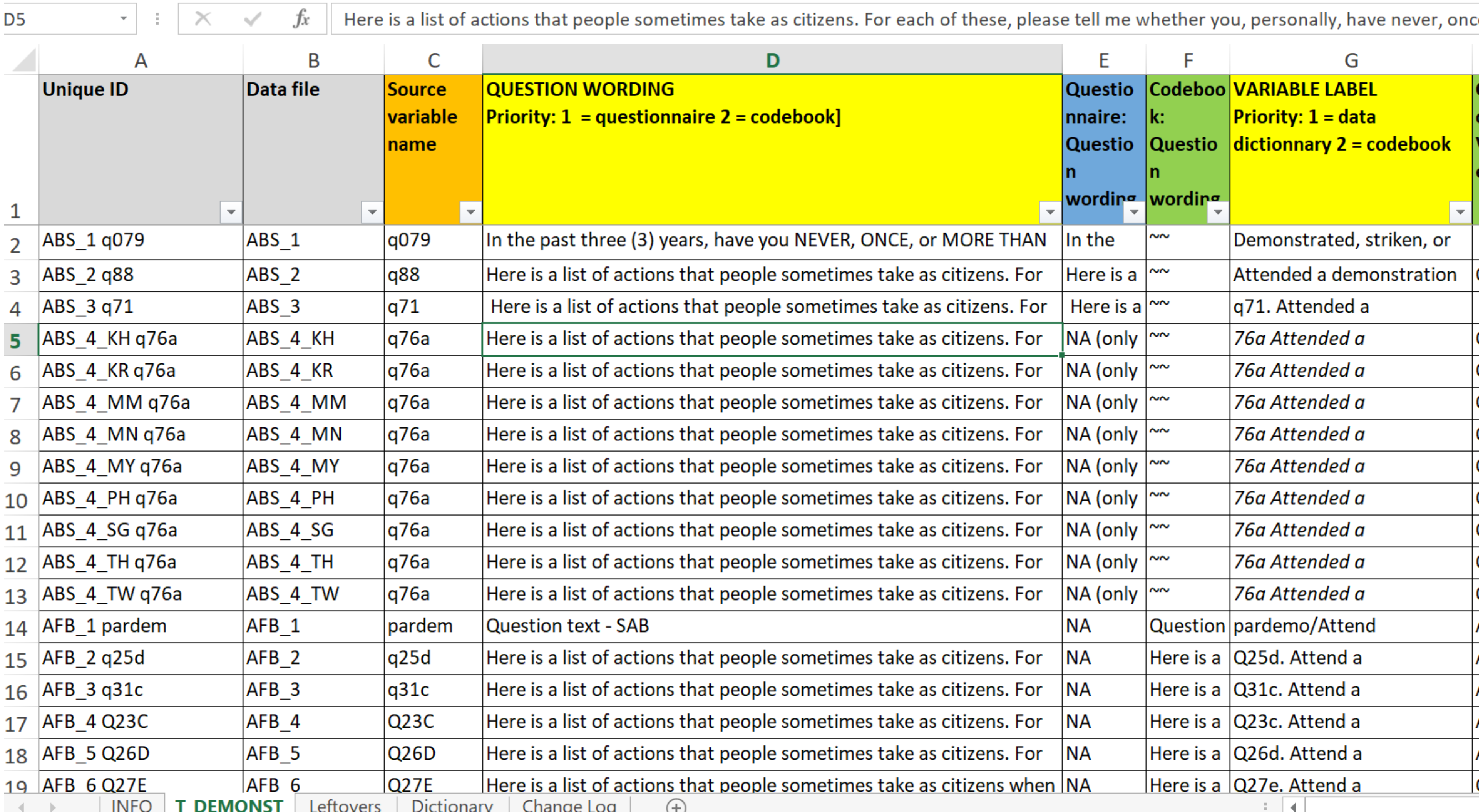
Work presented here reflects the collaboration of the SDR Team of faculty and graduate students from The Ohio State University and the Polish Academy of Sciences (asc.ohio-state.edu/dataharmonization/about/project-team/).

The *Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research*, SDR Project, is funded by the National Science Foundation, USA

(NSF SMA-1738502, Sept. 2017- Sept 2022).

SDR Harmonization Workflow

Detailed Variable Report DVR



D5 | fx | Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have never, onc

	A	B	C	D	E	F	G
1	**Unique ID**	**Data file**	**Source variable name**	**QUESTION WORDING Priority: 1 = questionnaire 2 = codebook]**	**Questionnaire: Question wording**	**Codebook: Question wording**	**VARIABLE LABEL Priority: 1 = data dictionary 2 = codebook**
2	ABS_1 q079	ABS_1	q079	In the past three (3) years, have you NEVER, ONCE, or MORE THAN	In the	~~	Demonstrated, striken, or
3	ABS_2 q88	ABS_2	q88	Here is a list of actions that people sometimes take as citizens. For	Here is a	~~	Attended a demonstration
4	ABS_3 q71	ABS_3	q71	Here is a list of actions that people sometimes take as citizens. For	Here is a	~~	q71. Attended a
5	ABS_4_KH q76a	ABS_4_KH	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
6	ABS_4_KR q76a	ABS_4_KR	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
7	ABS_4_MM q76a	ABS_4_MM	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
8	ABS_4_MN q76a	ABS_4_MN	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
9	ABS_4_MY q76a	ABS_4_MY	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
10	ABS_4_PH q76a	ABS_4_PH	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
11	ABS_4_SG q76a	ABS_4_SG	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
12	ABS_4_TH q76a	ABS_4_TH	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
13	ABS_4_TW q76a	ABS_4_TW	q76a	Here is a list of actions that people sometimes take as citizens. For	NA (only	~~	*76a Attended a*
14	AFB_1 pardem	AFB_1	pardem	Question text - SAB	NA	Question	pardemo/Attend
15	AFB_2 q25d	AFB_2	q25d	Here is a list of actions that people sometimes take as citizens. For	NA	Here is a	Q25d. Attend a
16	AFB_3 q31c	AFB_3	q31c	Here is a list of actions that people sometimes take as citizens. For	NA	Here is a	Q31c. Attend a
17	AFB_4 Q23C	AFB_4	Q23C	Here is a list of actions that people sometimes take as citizens. For	NA	Here is a	Q23c. Attend a
18	AFB_5 Q26D	AFB_5	Q26D	Here is a list of actions that people sometimes take as citizens. For	NA	Here is a	Q26d. Attend a
19	AFB_6 Q27E	AFB_6	Q27E	Here is a list of actions that people sometimes take as citizens when	NA	Here is a	Q27e. Attend a