Survey Data Recycling and Data Harmonization

Irina Tomescu-Dubrow, IFiS PAN and CONSI RT
Ilona Wysmulek, IFiS PAN

CONSIRT.osu.edu
Survey Data Harmonization

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

Data producers $\rightarrow$ Ex-ante/prospective harmonization
  • before fieldwork (input harmonization)
  • during data processing, before data release (output harmonization)

Secondary users $\rightarrow$ 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
**Integrated** SDR Database v.2.0

<table>
<thead>
<tr>
<th>Time span</th>
<th>1966-2017</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nr. countries</td>
<td>156</td>
</tr>
<tr>
<td>Nr. respondents</td>
<td>4,402,489</td>
</tr>
<tr>
<td>Nr. survey projects</td>
<td>23</td>
</tr>
<tr>
<td>Nr. project waves</td>
<td>174</td>
</tr>
<tr>
<td>Nr. national surveys</td>
<td>3,329</td>
</tr>
<tr>
<td>Nr. source data files</td>
<td>214</td>
</tr>
</tbody>
</table>

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
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

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*
Step four: CONTROL VARIABLES

SDR Harmonization Workflow

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
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: Mapping 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

<table>
<thead>
<tr>
<th>filter variable</th>
<th>item #</th>
<th>target label</th>
<th>prepare</th>
<th>make all SQL</th>
<th>parameter 1</th>
<th>check src value</th>
<th>clear</th>
<th>parameter 2</th>
<th>make SQL</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABS_4_KR q76a</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>miss</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>yes</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>would</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>would not</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>dk</td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_KR q76a</td>
<td>3</td>
<td>ref</td>
<td></td>
<td></td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>yes</td>
<td></td>
<td></td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>would</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>would not</td>
<td></td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ABS_4_MM q76a</td>
<td>1</td>
<td>miss</td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
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)
Step one: THEORY
- Time zero: data and documentation downloaded

Step two: SOURCE CHECK
- COTTON FILE (Excel, source data dictionary)

Step three: DEFINE TARGET
- MASTERFILE (Harmonized dataset)
- DATA USER GUIDE (i.a. General variable reports)

Step four: CONTROL VARIABLES
- DETAILED VARIABLE REPORT & QUALITY CONTROLS PLUG(s)

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)
39 substantive target variables (T), by color:

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

71 harmonization controls (arcs)
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/).

### SDR Harmonization Workflow

**Detailed Variable Report DVR**

Here is a list of actions that people sometimes take as citizens. For each of these, please tell me whether you, personally, have never, once, or more than once in the past three (3) years:

<table>
<thead>
<tr>
<th>Unique ID</th>
<th>Data file</th>
<th>Source variable name</th>
<th>QUESTION WORDING</th>
<th>Codebook: Questionnaire: Question wording Priority: 1 = questionnaire 2 = codebook dictionary 3 = codebook</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>ABS_1</td>
<td>q079</td>
<td>In the past three (3) years, have you NEVER, ONCE, or MORE THAN ONE TIME</td>
<td></td>
</tr>
<tr>
<td>A2</td>
<td>ABS_2</td>
<td>q88</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A3</td>
<td>ABS_3</td>
<td>q71</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A4</td>
<td>ABS_4</td>
<td>q76a</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A5</td>
<td>ABS_4</td>
<td>q76b</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A6</td>
<td>ABS_4</td>
<td>q76c</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A7</td>
<td>ABS_4</td>
<td>q76d</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A8</td>
<td>ABS_4</td>
<td>q76e</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A9</td>
<td>ABS_4</td>
<td>q76f</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A10</td>
<td>ABS_4</td>
<td>q76g</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A11</td>
<td>ABS_4</td>
<td>q76h</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A12</td>
<td>ABS_4</td>
<td>q76i</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A13</td>
<td>ABS_4</td>
<td>q76j</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A14</td>
<td>AFB_1</td>
<td>pardem</td>
<td>Question text - SAB</td>
<td></td>
</tr>
<tr>
<td>A15</td>
<td>AFB_2</td>
<td>q25d</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A16</td>
<td>AFB_3</td>
<td>q31c</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A17</td>
<td>AFB_4</td>
<td>Q3C</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A18</td>
<td>AFB_4</td>
<td>Q26D</td>
<td>Here is a list of actions that people sometimes take as citizens. For</td>
<td></td>
</tr>
<tr>
<td>A19</td>
<td>AFB_6</td>
<td>Q27E</td>
<td>Here is a list of actions that people sometimes take as citizens when</td>
<td></td>
</tr>
</tbody>
</table>