

# Data Harmonization in the Social Sciences: An Interdisciplinary Approach

Irina Tomescu-Dubrow, IFiS PAN, CONSIRT

Christof Wolf, GESIS

Kazimierz M. Slomczynski IFiS PAN, CONSIRT

J. Craig Jenkins, OSU

# Data Harmonization

Harmonization is about comparability

Data producers:

→ before fieldwork begins – Input harmonization

→ during data processing – Ex-ante output harmonization

Data users (+ some data producers):

→ after public data release – Ex-post output (retrospective) harmonization

## A Brief History of SDR

2013 – 2016 Democratic Values and Protest Behavior:  
Data Harmonization, Measurement Comparability, and Multi-Level  
Modeling in Cross-National Perspective (NCN 2012/06/M/HS6/00322)

Substantive interest → methodological implications →



Concern with feasibility

Academic background



decisions about specialized literature, data  
and measures

## Specialized literature

Substantive literature

Political sociology

Social psychology

Political science

Methodological literature

Survey research methodology (TSE, SQM)

Comparative survey research

Ex-post harmonization of survey data

Multi-level regression models

# The SDR Database V.1

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

22 projects; 89 waves; 1,721 national surveys; 142 countries/territories, btw. 1966 - 2013.  
N= 2, 289,060 respondents.

Available via Harvard dataverse ([asc.ohio-state.edu/dataharmonization/data/](https://asc.ohio-state.edu/dataharmonization/data/))

## Substantive target variables, SDR v.1.1

Political attitudes & behavior

Trust in the national parliament

Trust in the justice system

Trust in political parties

Trust in government

Participation in demonstrations

Signing petitions

Interest in politics

+

Trust in people (social trust)

Socio-demographics

Gender

Age

Year of birth

Rural/non-rural

Metropolitan

Education levels

Schooling years

## Methodological challenges

Large variation in data quality within and across international projects

Inter-survey differences in properties of specific items on the same issue

Increased total variability following harmonization partly due to non-random errors



## The survey data recycling (SDR) analytic framework (I)

Define, measure and store, as indicators, variability due to:

- source survey quality
- data reprocessing (i.e. ex-post harmonization)

Methodological biases and errors understood as consequences of:

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

## The SDR analytic framework (II)

Metadata stored as variables capture:

- (i) Unequal source data quality as reflected in:
  - (a) data records in computer files;
  - (b) survey documentation (codebook, questionnaires, technical reports, etc.)
  - (c) (in)consistency documentation – data records (i.e. processing errors)
  
- (ii) Variability stemming from ex-post harmonization
  - properties of scales
  - question wording
  - questionnaire context
  - deriving information for one target variable from “extra” source data

Researchers can use metadata variables as controls.

Work ongoing, SDR Project, 2017 -2021

## Co-edited volume on Survey data harmonization in the social sciences

Ex-ante and ex-post harmonization matter for comparability also in relation to each other

Rapidly growing number of harmonized survey datasets (ex-ante, ex-post); weak documentation of harmonization decisions (Granda and Blaszczyk 2016)

Small methodological literature on harmonization *per se*

Limited scope of *Harmonization*: Newsletter on Survey Data Harmonization in the Social Sciences

([asc.ohio-state.edu/dataharmonization/harmonization-newsletter-fall-2019/](http://asc.ohio-state.edu/dataharmonization/harmonization-newsletter-fall-2019/))

→ Proposed volume + Warsaw Event

# Main Themes of Data Harmonization - I

## A. Relevance for the survey life cycle

The Effects of Data Harmonization on the Survey Research Process, by Peter Granda, University of Michigan

## B. Bursting the disciplinary bubble

Harmonizing Census Data: IPUMS International, by Matthew Sobek

Harmonizing time-use data: current practices and challenges in the field, by Ewa Jarosz, Sarah Flood, Margarita Vega Rapun

Effecting rigorous data harmonization and documentation to understand data heterogeneity and quality, by Isabel Fortier and Tina W. Wey

## Main Themes of Data Harmonization - II

### C. Theory-informed ex-post harmonization

Harmonization of Panel Data: CNEF by Dean Lillard

Theory Driven Approach to Data Harmonization: The True European Voter Project, by Paolo Segatti and colleagues

SDR-based Roundtable discussions on Wednesday (Day 1 of Workshop)

### D. Inter-survey comparability of concepts

Using Latent Class Analysis for Testing the Equivalence of Multi-Item Scales across Data from Different Survey Programs. The Example of National Identity Types, by Markus Quandt and Antonia May

## Main Themes of Data Harmonization - III

### E. Data quality of source surveys and the data harmonized ex-post

Comparative Sample Quality: Facing the Obstacles to Applying External and Internal Evaluation Criteria to Large Harmonized Datasets,  
by Piotr Jabkowski and Piotr Cichocki

Effecting rigorous data harmonization and documentation to understand data heterogeneity and quality, by Isabel Fortier and Tina W. Wey

Inter-survey variability in source data quality, Workshop, Day 1, by SDR Team

Harmonization control variables, Workshop, Day 1, by SDR Team

## Main Themes of Data Harmonization - IV

### F. Creation and use of metadata

Harmonizing Census Data: IPUMS International, by Matthew Sobek

SDR as an analytic framework for survey data reprocessing, Workshop Day 1, by SDR Team

### G. Documentation of the ex-post harmonization process

From Source to Target: Harmonization, Workflow, Procedures and Tools, Workshop Day 1, by SDR Team

## Main Themes of Data Harmonization - V

### H. Statistical issues in analyzing harmonized data

Statistical Issues in Analyzing Harmonized Data, by Claire Durand

Applying Multiple Imputation to Survey Data Harmonized Ex-Post,  
Workshop, Day 2, by Stef van Buuren

### I. Dissemination of ex-post harmonized data; sustainability

HUMAN Surveys: Challenges of Harmonizing Multiple Sources Using Stata Scripts  
by Andrew Klassen



## Main Themes of Data Harmonization - VI

### J. The potential for new methodological and substantive knowledge

Analyzing repeated-wave panel data to identify causal direction: Lessons from a meta-analysis of digital media use and political participation, by Jennifer Oser

Three methodological papers in political sociology, on aggregation of survey data into measures of country characteristics – Workshop, Day 3, by Joonghyun Kwak & Craig Jenkins; Joonghyun Kwak & Kazimierz Slomczynski; and Joshua Dubrow & Olga Zelinska

Complexity in Society: From Indicators Construction to their Synthesis, Workshop, Day 3, by Filomena Maggino

## Acknowledgements

Work presented here, and this event itself, exist through collaboration. Thank you!

This event is organized by CONSIRT, The Ohio State University and the Polish Academy of Sciences.

Financial support comes from the US National Science Foundation (SMA-1738502), the Polish National Science Centre (NCN 2016/23/B/HS6/03916), as well as the Polish Academy of Sciences, and the Institute of Philosophy and Sociology PAN.