Re-Thinking *Ex Post* Harmonization for Comparative Research: Data Recycling of 22 International Survey Projects

Kazimierz M. Slomczynski, CONSIRT
Irina Tomescu-Durbow, CONSIRT
J. Craig Jenkins, The Ohio State University (OSU)

Cross-national Studies, Interdisciplinary Research and Training Program - CONSIRT, the Polish Academy of Sciences and OSU
consirt.osu.edu

Research supported by the Polish National Science Centre (2012/06/M/HS6/00322) as part of the study *Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling* (hereafter, Harmonization Project)
Criteria of selecting survey projects

Projects:
- containing questions about political attitudes and behaviors
- designed as cross-national, and, preferably, multi-wave;
- with the samples intended as representative of the adult population of given country or territory;
- non-commercial;
- freely available in the public domain;
- with documentation - study description, codebook and/or questionnaire - in English
<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Survey Project</th>
<th>Time span</th>
<th>Waves</th>
<th>Files</th>
<th>Data Sets</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFB</td>
<td>Afrobarometer</td>
<td>1999-2009</td>
<td>4</td>
<td>4</td>
<td>66</td>
<td>98942</td>
</tr>
<tr>
<td>AMB</td>
<td>Americas Barometer</td>
<td>2004-2012</td>
<td>5</td>
<td>1</td>
<td>92</td>
<td>151341</td>
</tr>
<tr>
<td>ARB</td>
<td>Arab Barometer</td>
<td>2006-2011</td>
<td>2</td>
<td>2</td>
<td>16</td>
<td>19684</td>
</tr>
<tr>
<td>ASB</td>
<td>Asian Barometer</td>
<td>2001-2011</td>
<td>3</td>
<td>3</td>
<td>30</td>
<td>43691</td>
</tr>
<tr>
<td>ASES</td>
<td>Asia Europe Survey</td>
<td>2000</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td>18253</td>
</tr>
<tr>
<td>CB</td>
<td>Caucasus Barometer</td>
<td>2009-2012</td>
<td>4</td>
<td>4</td>
<td>12</td>
<td>24621</td>
</tr>
<tr>
<td>CDCEE</td>
<td>Consolidation of Democracy (C./East Europe)</td>
<td>1990-2001</td>
<td>2</td>
<td>1</td>
<td>27</td>
<td>28926</td>
</tr>
<tr>
<td>CNEP</td>
<td>Comparative National Elections Project</td>
<td>2004-2006</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>13372</td>
</tr>
<tr>
<td>EB</td>
<td>Eurobarometer</td>
<td>1983-2012</td>
<td>7</td>
<td>7</td>
<td>152</td>
<td>138753</td>
</tr>
<tr>
<td>EQLS</td>
<td>European Quality of Life Survey</td>
<td>2003-2012</td>
<td>3</td>
<td>1</td>
<td>93</td>
<td>105527</td>
</tr>
<tr>
<td>ESS</td>
<td>European Social Survey</td>
<td>2002-2013</td>
<td>6</td>
<td>2</td>
<td>146</td>
<td>281496</td>
</tr>
<tr>
<td>EVS/WVS</td>
<td>European Values Study/World Values Survey</td>
<td>1981-2009</td>
<td>9</td>
<td>1</td>
<td>312</td>
<td>423084</td>
</tr>
<tr>
<td>ISJP</td>
<td>International Social Justice Project</td>
<td>1991-1996</td>
<td>2</td>
<td>1</td>
<td>21</td>
<td>25805</td>
</tr>
<tr>
<td>ISSP</td>
<td>International Social Survey Programme</td>
<td>1985-2013</td>
<td>13</td>
<td>13</td>
<td>363</td>
<td>493243</td>
</tr>
<tr>
<td>LB</td>
<td>Latinobarometro</td>
<td>1995-2010</td>
<td>15</td>
<td>15</td>
<td>260</td>
<td>294965</td>
</tr>
<tr>
<td>LITS</td>
<td>Life in Transition Survey</td>
<td>2006-2010</td>
<td>2</td>
<td>2</td>
<td>64</td>
<td>67866</td>
</tr>
<tr>
<td>NBB</td>
<td>New Baltic Barometer</td>
<td>1993-2004</td>
<td>6</td>
<td>1</td>
<td>18</td>
<td>21601</td>
</tr>
<tr>
<td>PA2</td>
<td>Political Action II</td>
<td>1979-1981</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>4057</td>
</tr>
<tr>
<td>PA8NS</td>
<td>Political Action – 8 Nation Study</td>
<td>1973-1976</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>12588</td>
</tr>
<tr>
<td>PPE7N</td>
<td>Political Participation and Equality in 7 Nations</td>
<td>1966-1971</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>16522</td>
</tr>
<tr>
<td>VPCPCE</td>
<td>Values/Political Change, Postcommunist E</td>
<td>1993</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>4723</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1966-2013</td>
<td>89</td>
<td>81</td>
<td>1721</td>
<td>2289060</td>
</tr>
</tbody>
</table>
Data

• 1,721 national surveys with over 95% of samples from 500 to 3,000 respondents

• National surveys conducted in 142 countries and territories over a period of almost 50 years

• All these surveys contain over 25,000 variables

• From 150 to 200 variables are identical or similar in large majority of 89 waves

• From 25 to 40 variables deal directly with political attitudes and behaviors
Survey Data Recycling (SDR)

- SDR = framework for (re)-processing cross-national survey data;
- SDR concerns survey data quality control & *ex-post* survey harmonization, to:
  (a) account for “messiness” of the original source data,
  (b) provide “comparable” data
Thus, SDR expands the scope of extant projects (time, space, number of observations, types of indicators)

**Unifying thesis:** account for errors & biases in original surveys & harmonization procedures *via* different types of quality control variables, to be included in substantive analyses.
Figure 1. General Schema of Survey Data Recycling

Input
Materials on national surveys from different sources and in different forms

Organized original materials
- General survey documentation
- Specific description of the data
- Computer data files

Transformed materials
- Standardized survey documentation
- Integrated codebook
- Data file of selected source variables

Quality evaluation
- Quality-control variables for general survey documentation, specific data description, & the computer data file
- Weights

Ex-post Harmonized data
- “New” technical variables
- Substantive target variables
  - Quality-control variables for target variables
  - Cross-national comparability checks

Output
INTEGRATED DATABASE
Standardized documentation and integrated codebook; integrated data file with harmonized technical & substantive variables, quality control indicators
Formal framework

1. Relationship between target variables T and source variables S
   \[ T = f(S) \quad \text{[substantive decision of } f \text{]} \]

2. Relationship between T and X and two types of control variables: Q and H
   \[ X = \text{Substantive independent variables} \]
   \[ Q = \text{Data Quality Controls for (a) survey documentation, (b) specific data description, and (c) data in the computer files} \]
   \[ H = \text{Harmonization Quality Controls of specific procedures that could influence validity and reliability of } T \]
   \[ T = b_0 + b_1 Q + b_2 H + b_3 X + e \]
Explanation

T = Target Variable
Q = Data Quality Controls
H = Harmonization Quality Controls
X = Substantive Independent Variables

\[ T = b_0 + b_1 Q + b_2 H + b_3 X + e \]

If \( e \) is negligible & \( b_1, b_2, b_3 = 0, b_0 = f(S) \) for \( T = f(S) \)
If \( b_1 \) and/or \( b_2 \) unequal 0, some intervention may be needed to correct for errors and biases in \( T \).

Minimal solution: partialing out effects of \( Q \) and \( H \) in the estimated impact of \( X \) on \( T \).
Survey quality-control Indicators

(a) survey documentation;

(b) (in)consistency between data description (e.g. in codebooks, questionnaires) & data records and data records in computer file;

(c) computer data records themselves.
General Survey Documentation: How the data has been collected?

<table>
<thead>
<tr>
<th>Question</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the survey documentation specify the type of sample used?</td>
<td>Yes = 0 No = 1</td>
</tr>
<tr>
<td>Does the survey documentation provide information on the response rate?</td>
<td>Yes = 0 No = 1</td>
</tr>
<tr>
<td>Was the questionnaire back-translated or translation checked in some other way?</td>
<td>Yes = 0 No = 1</td>
</tr>
<tr>
<td>Is there any evidence that the questionnaire was pre-tested?</td>
<td>Yes = 0 No = 1</td>
</tr>
<tr>
<td>Does the documentation show that the fieldwork was controlled?</td>
<td>Yes = 0 No = 1</td>
</tr>
</tbody>
</table>

Effect of negative answers (No = 1): Reduction of confidence in the data
### Specific Data Description: How have the data been defined?

<table>
<thead>
<tr>
<th>Question</th>
<th>Answers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Do variable values in the codebook correspond to values in the data file?</td>
<td>Yes = 0, No = 1</td>
</tr>
</tbody>
</table>

**Eight binary variables describing discrepancies between data description and the data file**

(Ilona and Olena)

**Effect of negative answers (No = 1): Decrease of interpretability of the data**
<table>
<thead>
<tr>
<th>Computer Data File: Are the data formally correct?</th>
<th>Answers</th>
</tr>
</thead>
</table>
| Do survey cases (respondents) have unique identification numbers (IDs)? | Yes = 0  
No = 1 |
| Are survey weights free of formal errors? (Marcin and Przemek) | Yes = 0  
No = 1 |
| Is the proportion of missing values for gender and age within the standard limits (< 5%)? | Yes = 0  
No = 1 |
| Is the data file free from repeated cases (duplicates)? | Yes = 0  
No = 1 |
| Effect of negative answers (No = 1) : Possible distortion of the research results based on the data |         |
Data Harmonization Controls

Variables pertaining to:

• Wording

• Scales

• Item Context

[Measurement properties]

(Marta paper on wording, scales, and item context)
Harmonization Project Team

• Kazimierz M. Słomczyński, Polish Academy of Sciences and Ohio State University

• Irina Tomescu-Dubrow, CONSIRT

• Craig Jenkins, Ohio State University

• Marta Kołczyńska, Ohio State University

• Ilona Wysmułek, Graduate School for Social Research (Warsaw)

• Olena Oleksiyenko, Graduate School for Social Research (Warsaw)

• Przemek Powałko, Polish Academy of Sciences

• Marcin Zieliński, University of Warsaw

• Matt Schoene, Ohio State University

and

Joshua Dubrow, Anna Franczak, Anastas Vangeli, Jakub Wysmulek - Graduate School for Social Research (Warsaw)