Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling

Irina Tomescu-Dubrow & Kazimierz M. Slomczynski


To link to this article: http://dx.doi.org/10.1080/00207659.2016.1130424

Published online: 08 Mar 2016.

Article views: 10

Submit your article to this journal

View related articles

View Crossmark data
Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling

Irina Tomescu-Dubrow and Kazimierz M. Slomczynski

Institute of Philosophy and Sociology, Polish Academy of Sciences

This article describes challenges and solutions to ex post harmonization of survey data in the social sciences based on the big data project “Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling.” This project engages with the relationship between democracy and protest behavior in comparative perspective by proposing a theoretical model that explains variation in political protest through individual-level characteristics, country-level determinants, and interactions between the two. Testing it requires data with information at both the individual and country levels that vary across space and over time. The project’s team pooled information from 22 well-known international survey projects into a data set of 2.3 million respondents, covering a total of 142 countries and territories, and spanning almost 50 years, to construct common measures of political behavior, social attitudes, and demographics. The integrated data set is appended with country variables from nonsurvey sources. Mapping the methodological complexities this work raised and their solutions became the springboard for the analytic framework of Survey Data Recycling (SDR). SDR facilitates reprocessing information from extant cross-national projects in ways that minimize the “messiness” of data built into original surveys, expand the range of possible comparisons over time and across countries, and improve confidence in substantive results.

Keywords cross-national; data harmonization; survey data recycling

The project “Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling” (Harmonization Project, hereafter) is driven by

Irina Tomescu-Dubrow is an associate professor of sociology at the Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN), and program manager for the Cross-National Studies: Interdisciplinary Research and Training program of Ohio State University and PAN. She studies social inequality, stratification, and structural change, and is involved in cross-national methodology research, including ex post survey data harmonization. Her research has appeared in Quality and Quantity, Journal of Urban Affairs, Polish Sociological Review, the International Journal of Sociology, Problems of Post-Communism, and Ask: Research and Methods.

Kazimierz M. Slomczynski is a professor at the Institute of Philosophy and Sociology at the Polish Academy of Sciences (PAN). He is also director of the Cross-National Studies: Interdisciplinary Research and Training program of Ohio State University and PAN. He is principal investigator of the Polish Panel Survey POLPAN, conducted every five years since 1988, and directs other studies, including a project on ex post harmonization of cross-national surveys. He has published numerous articles in various international scientific journals and authored or coauthored several books in Polish and English. His main interest is in social stratification and methodology of social sciences.

Address correspondence to Irina Tomescu-Dubrow, Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN), Nowy Swiat 72, 00-330 Warsaw, Poland. E-mail: tomescu.1@osu.edu.
interrelated substantive and methodological considerations. Substantively, the project engages with the relationship between democracy and protest behavior in comparative, cross-national perspective. We focus on two different types of conventional political protests: participation in demonstrations and signing petitions. Regarding democracy, we consider both democratic practice—usually measured by “democratic indexes” characterizing countries—and people’s attitudes toward the political system and its institutions. The Harmonization Project incorporates relevant literature to develop a theoretical model that explains variation in political protest in light of individual-level characteristics, country-level determinants, and interactions between the two types of factors. Testing it calls for data at the individual and country levels that vary over time and across space.

The social sciences have a growing wealth of cross-national surveys, yet the data are often not directly comparable across different surveys. Questions about participation in demonstrations and signing petitions are asked differently in major projects such as the World Values Survey, European Social Survey (ESS), or International Social Survey Programme (ISSP). In addition, major international projects, such as the European Social Survey, Latino Barometer, and Asia-Europe Survey, among others, are region-oriented, which hinders research pertaining to world-relevant issues. To address these challenges as they appear in relation to its substantive research problem, the Harmonization Project set out to create comparable measurements of political protest, social values, and demographics via ex post harmonization of variables from international survey projects, and append them with macro-level variables from external sources, such as the World Bank, Organization for Security and Co-operation in Europe (OSCE), United Nations (UN) agencies, Transparency International, and others (dataharmonization.org/data).

Using data from 1,721 national surveys, the harmonization team created an integrated data set for 142 countries and territories. Figure 1 shows that it covers well most of Europe and large

![FIGURE 1 Geographic coverage on national surveys in the Harmonization Project.](image-url)
parts of the Americas and Asia, but has noticeable gaps, especially in Africa (including all of Central Africa) and Oceania (for a more detailed discussion of coverage, see Tomescu-Dubrow and Slomczynski 2014).

To construct common measurements across the pooled data required many preparatory steps, including evaluating and standardizing the general survey documentation, questionnaires, and the computer data files. It was necessary to develop and document the methodology for each of them. Next, the harmonization process called for decisions on, among others, the function relating target and source variables, the adequacy of given computational or statistical tools for implementing harmonization rules, or the validity and reliability of target variables. Together, these considerations prompted our work toward a comprehensive theoretical and methodological base for ex post harmonization of survey data, under the new framework of Survey Data Recycling (SDR) (Slomczynski and Tomescu-Dubrow 2015).

**THE HARMONIZATION PROJECT: SUBSTANTIVE AND METHODOLOGICAL CONSIDERATIONS**

Current studies on protest behavior in many, mainly European, countries focus on micro determinants, such as gender, age, education, and interest in politics. As Gallego (2007) points out, people’s socioeconomic and demographic characteristics are related to their acquisition of resources, which in turn influence the costs of participating in politics. The relevance of people’s interest in politics for both electoral and nonelectoral political action is also well-established in the literature (e.g., Leighley and Vedlitz 1999; Rosenstone and Hansen 1993; Tomescu-Dubrow and Slomczynski 2014).

Limiting explanatory models to individual characteristics is not justified on empirical grounds because over the world there is “marked variation in protest across nations, with a 20:1 ratio in protest mean-scores between the highest-ranking (Sweden) and lowest-ranking (Vietnam) nations” (Dalton, Sickle, and Weldon 2009: 14). In recent years researchers have paid increasing attention to balancing micro- and macro determinants of political participation (Benson and Rochon 2004; Dubrow, Slomczynski, and Tomescu-Dubrow 2008; Kriesi 2004; Marien, Hooghe, and Quintelier 2010; Vrablikova 2013). The Harmonization Project takes a similar stance.

From the standpoint of methodology, research on protest is generally based on some particular data sets involving limited coverage of countries and topics; various kinds of protest activities are inconsistently combined, and likely produce biased measures; the statistical techniques employed in analyses are not fully satisfactory because they often fail to deal explicitly with the hierarchical structure of the data (i.e., people nested in countries and time intervals).

In response to such problems, the Harmonization Project (a) proposed a multilevel model of political protest that includes characteristics of people and of countries, and cross-level interactions, and (b) created, through reprocessing information from extant cross-national survey projects, a new data set with broader time and space coverage than is found in any existing single project, and appended it with relevant country-level variables. We will review both solutions after discussing harmonization of social science survey data, which is foundational to this project.
Cross-National Survey Data Harmonization in the Social Sciences

Data harmonization is a generic term for procedures that aim to achieve, or at least improve, the comparability of surveys over time and of surveys from different countries (Granda and Blasczyk 2010; Granda, Wolf, and Hadorn 2010). As a methodological framework, it is especially relevant to studying relationships between any two major concepts in cross-national perspective because it facilitates greater between-country variation. (For a detailed discussion of the rise of cross-national survey data harmonization in the social sciences, see Dubrow and Tomescu-Dubrow [2015]).

Extant international survey projects encounter shortcomings such as regional focus (e.g., the European Social Survey, the various barometer studies), underrepresentation of historically marginalized regions, and limited country coverage (even in Europe, as documented in Slomczynski and Tomescu-Dubrow 2006), or restrictions on the measurement of major concepts (e.g., the World Values Survey). To overcome them, and to increase the scope of their studies, researchers often run analyses on several surveys, but separately. Cross-national survey data harmonization proposes a different solution: it strives to pool and adjust (i.e., recode, rescale, transform) different survey data sets not a priori designed to be compared, into a new, integrated, data set that could be analyzed as a typical single data source. The literature refers to this process as ex post harmonization, to the original variables in the data sets of particular surveys as source variables, and to the harmonized, common, variable produced from the source variables as target variables (Ehling, Rendtel et al. 2006; Granda and Blasczyk 2010; Granda et al. 2010; Gunther 2003).

Ex post survey harmonization is rife with methodological challenges, and the Harmonization Project is no exception. These include dealing with various types of errors in the original data (for details on the Total Survey Error paradigm, see Biemer [2010] and Smith [2011]). Because the soundness of empirical statements based on survey data depends on their quality, concerns with the data records themselves are paired with concerns about the quality of the source documentation and the consistency of documentation with data records (Slomczynski and Tomescu-Dubrow 2014).

Of paramount importance to ex post harmonization is cross-national measurement equivalence. It is beyond the scope of this article to engage with the wealth of literature on this topic (see, e.g., Byrne and van de Vijver 2010; Cheung 2008; Harkness 1998; Harkness, van de Vijver, and Mohler 2003; Jowell 1998; Jowell et al. 2007; Kennett and Yeates 2001; Matsumoto and van de Vijver 2010; Medina, Smith, and Long 2009; Przeworski and Teune 1972). Instead, we draw on this accumulated knowledge and on our experience in the Harmonization Project to propose that target variables also be evaluated through quality indicators of specific harmonization procedures that could influence their validity and reliability.

To illustrate the need for controlling face validity of the constructs, we refer to different formulations of the questionnaire items on participation in demonstrations. A straightforward question is: “During the last 12 months, have you done any of the following? Have you taken part in a lawful public demonstration?” coded 1 = Yes, 2 = No, 7 = Refusal, 8 = Don’t know (European Social Survey). The World Values Survey uses more than one measure when asking individuals about their protest behavior, including a dichotomous Yes/No, but adds a temporal component to it about whether or not the respondent has engaged in the behavior in the past three years or in the past five years. In other international survey projects, the similar question contains the word “demonstrations” together with marches and rallies; some refer to legal and illegal
action. In sum, with respect to “demonstration,” the questionnaire items in the major international survey projects differ with respect to the time frame in which engagement in this kind of behavior occurs, the number of items related to the concept, whether additional terms were used besides “demonstration,” and whether there is reference to illegality of the action. The solution we propose in the Survey Data Recycling framework is to construct control variables describing peculiarity of the item in each national survey, and to include all control variables in substantive analyses. This does not replace, but rather complements cross-national comparability checks based on intergroup comparisons of confirmatory factor analyses, or other statistical techniques, as applicable. We will return to the issue of quality control indicators later in the article.

THEORETICAL MODEL OF POLITICAL PROTEST AND ITS DETERMINANTS

The Harmonization Project proposed a multilevel model that explains political protest (individual-level) by a set of theoretically informed micro-level and macro-level (contextual) variables. People engage in protest to various degrees, depending on personal characteristics—civic skills, economic circumstances, gender, age, and ideology—to mention a few. Yet these individual-level conditions for action are themselves shaped by the context in which we live. We consider two types of conventional protest (Jenkins and Form 2005). The first is participation in public demonstrations, a good example of collectivistic behavior in that “the act is designed as [a] physical display of opinion to build solidarity, and the person has the sense of being part of a gathering” (Tomescu-Dubrow and Slomczynski 2014: 136). The second form of protest, signing a petition, contains both collectivistic (preparing a petition calls for cooperation) and individualistic aspects (petition signing is a personal act) (Dubrow et al. 2008).

Figure 2 depicts the expected relationships. The leading hypothesis is that participation in political protest, $Y$, is a linear function of a set of sociodemographic variables ($X_1$), political participation other than protest ($X_2$), trust in public institutions ($X_3$), and democratic values ($X_4$)—measured at the respondent level—as well as economic development ($W_1$) and democratic practices ($W_2$)—measured at the country level. The schema also depicts the discrepancy between countries’ democratic practices and people’s democratic values, expressed as a cross-level interaction ($W_2 \times X_4 = X_5$).

Most of these constructs and the relations between them are well discussed in the literature, often in light of competing theories: relative deprivation theory (Auvinen 1997; Gurr and Duval 1973; Lichbach 1989); resource mobilization (Inglehart 1990; Jenkins, Wallace, and Fullerton 2008; McAdam, McCarthy, and Zald 1996); political values/ideologies (Inglehart and Catterberg 2002; Inglehart and Welzel 2005); political opportunity structure (Kitschelt 1986; Kriesi 2004; Meyer and Minkoff 2004); or the individual–institutional nexus of protest behavior (Dalton et al. 2009).

Two observations are in place. First, country-level indicators of democracy from nonsurvey sources allow one to measure straightforwardly discrepancies between institutional democratic performance on one hand, and people’s attitudes toward/values on democracy, on the other via a cross-level interaction term. We expect such tensions to matter for the extent and intensity of political protest, and refer to them explicitly because the literature does not cover them. Other cross-level interactions, such as those Dalton and his colleagues (2009) discuss in the individual–institutional nexus of protest behavior are also of interest to us. Second, at the
country level, measures of ethnic fractionalization, economic development, and corruption are also worth considering. In short, it is clear that additional data from nonsurvey sources are intrinsic to this type of analysis.

The generic, simplified equation for the multilevel model, for a given time \( t \), is as follows (Tomescu-Dubrow and Slomczynski 2014: 105):

\[
Y = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + u_{1j}X_{ij} + u_{0j} + e_{ij},
\]

where \( Y \), the dependent variable, varies among individuals \( i \) and countries \( j \). \( X \) is a vector of individual characteristics and \( W \) is a vector of macro-level characteristics. Note that:

- \( \gamma_{00}, \gamma_{10}, \text{ and } \gamma_{01} \) are fixed coefficients/effects
- \( u_{ij}, u_{0j}, e_{ij} \) are random coefficients/effects.

In this study, we define key terms for multilevel modeling in the cross-national context, beginning with random effects and fixed effects. Random effects, which are estimated as variance components, are model parameters that are estimated to vary between countries, whereas fixed effects are estimates that are constant across countries. We have the random error term \( \mu_{0j} \), which is the part of the model that represents the intercept variance between countries. In addition to this error term, we have \( \mu_{1ij} \), which allows the random slope for individual-level
predictor variable $X_{ij}$. A random slope means that the relationship between the individual-level predictor ($X_{ij}$) and the outcome ($Y_{ij}$) are allowed to vary across countries.

In Equation (1), $\gamma_{00}$, $\gamma_{10}$, and $\gamma_{01}$ are analogous to unstandardized regression coefficients in a single-level model. More specifically, the fixed effects represent: (a) the model intercept ($\gamma_{00}$), which represents the grand mean of our dependent variable—some protest behavior scale $Y$—across individuals and countries; (b) the fixed effect for $X_{ij}$ ($\gamma_{10}$), which represents the relationship between a given predictor on the individual level and the dependent variable while controlling for all other variables in the model; and (c) the fixed effect for $W_j$ ($\gamma_{01}$), which represents the relationship between country-level variable and political participation while controlling for all other variables in the model. If we assume that the effect of the predictor $X$ on $Y$ depends on the country-level variable $W$, the interaction term $X_{ij}W_j$ must be introduced, resulting in the equation:

$$Y = \gamma_{00} + \gamma_{10}X_{ij} + \gamma_{01}W_j + \gamma_{11}X_{ij}W_j + u_{1j}X_{ij} + u_{0j} + e_{ij}.$$  

(2)

Under these specifications, in Equations (1) and (2) error terms are heteroskedastic instead of homoskedastic, as it is assumed in ordinary regression models where the residual errors are considered independent of the values of the explanatory variable. Dealing with the problem of heteroskedasticity is one of the main reasons for preferring multilevel models over the regular ordinary least squares (OLS) models when analyzing hierarchical nested data (see Gelman and Hill 2007; Hox 2010).

Extending Equations (1) and (2) for varying time $t$ leads to more complicated models, which would involve interaction terms of the individual-level variables with country-specific time-varying variables. To further complicate matters, we could include time-varying variables that are not country-specific. In this case, the models would be nonhierarchical multilevel models.

**INTEGRATED DATA SET IN THE HARMONIZATION PROJECT**

To increase geographic and temporal variability in the data, as is necessary to analyze the relation between democracy and political protest in comparative perspective, the harmonization team selected 22 well-known international survey projects, listed in Table 1, which meet the following criteria: they are noncommercial; designed as cross-national, and, preferably, multi-wave; the samples are intended to be representative of the adult population of a given country or territory; they contain questions about political attitudes and behaviors; they are freely available in the public domain; their documentation—study description, codebook and/or questionnaire—is provided in English. In all survey projects, the units of observation are individuals.

From the 22 projects, we selected 89 waves (i.e., project*wave) that are relevant for political protest. Using Structured Query Language (SQL), we constructed a relational database that contains 81 data files with 1,721 project*wave*countries (i.e., national surveys carried out in all waves of all projects) and a total of almost 2.3 million respondents. A single file may contain data from one country in one wave to many countries in many waves, hence the difference between number of waves (89) and the number of data files (81). The platform for data files of national surveys is organized such that in the future, any variable could be extracted and moved to the integrated data set (Powało 2015; and Powało and Kołczyńska in this issue).
The relational database stores data in different types of tables (see Powalko and Kołczyńska in this issue). Among them, source data tables and source metadata tables produce the harmonized master table (i.e., the integrated table). The master table includes four types of variables: (1) standardized technical variables; (2) source variables, preserved for reference; (3) target variables resulting from harmonization of source variables; and (4) harmonization quality controls. These data are matched with control indicators for the quality of the source data and with country-level data. We are in the process of selecting from electronic and published sources (e.g., World Bank Indicators, UN Data [data.un.org], the Quality of Governance database [qog.pol.gu.se]) country-level structural variables, including types of political systems, indexes of democracy, corruption, gross domestic product (GDP), national unemployment rates, and measures of income inequality. Information from all tables in the relational database can be exported to a CSV text file, and further converted to any file format used in statistical packages.

### Table 1: Selected International Survey Projects

<table>
<thead>
<tr>
<th>Abbrev.</th>
<th>Survey project</th>
<th>Time span</th>
<th>Waves</th>
<th>Countries / territories</th>
<th>Files</th>
<th>Data sets&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFB</td>
<td>Afrobarometer</td>
<td>1999–2009</td>
<td>4</td>
<td>13</td>
<td>4</td>
<td>66</td>
<td>98,942</td>
</tr>
<tr>
<td>AMB</td>
<td>Americas Barometer</td>
<td>2004–2012</td>
<td>5</td>
<td>20</td>
<td>1</td>
<td>92</td>
<td>151,341</td>
</tr>
<tr>
<td>ARB</td>
<td>Arab Barometer</td>
<td>2006–2011</td>
<td>2</td>
<td>24</td>
<td>2</td>
<td>16</td>
<td>19,684</td>
</tr>
<tr>
<td>ASB</td>
<td>Asian Barometer</td>
<td>2001–2011</td>
<td>3</td>
<td>11</td>
<td>3</td>
<td>30</td>
<td>43,691</td>
</tr>
<tr>
<td>ASES</td>
<td>Asia-Europe Survey</td>
<td>2000</td>
<td>1</td>
<td>18</td>
<td>1</td>
<td>18</td>
<td>18,253</td>
</tr>
<tr>
<td>CB</td>
<td>Caucasus Barometer</td>
<td>2009–2012</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>12</td>
<td>24,621</td>
</tr>
<tr>
<td>CDCEE</td>
<td>Consolidation of Democracy&lt;sup&gt;b&lt;/sup&gt;</td>
<td>1990–2001</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>27</td>
<td>28,926</td>
</tr>
<tr>
<td>CNEP</td>
<td>Comparative National Elections Project&lt;sup&gt;c&lt;/sup&gt;</td>
<td>2004–2006</td>
<td>1</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>13,372</td>
</tr>
<tr>
<td>EB</td>
<td>Eurobarometer&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1983–2012</td>
<td>7</td>
<td>37</td>
<td>7</td>
<td>152</td>
<td>138,753</td>
</tr>
<tr>
<td>EQLS</td>
<td>European Quality of Life Survey</td>
<td>2003–2012</td>
<td>3</td>
<td>35</td>
<td>1</td>
<td>93</td>
<td>105,527</td>
</tr>
<tr>
<td>ESS</td>
<td>European Social Survey</td>
<td>2002–2013</td>
<td>6</td>
<td>32</td>
<td>2</td>
<td>146</td>
<td>281,496</td>
</tr>
<tr>
<td>EVS</td>
<td>European Values Study</td>
<td>1981–2009</td>
<td>4</td>
<td>50</td>
<td>1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>128</td>
<td>166,502</td>
</tr>
<tr>
<td>ISSP</td>
<td>International Social Survey Programme&lt;sup&gt;c&lt;/sup&gt;</td>
<td>1985–2013</td>
<td>13</td>
<td>53</td>
<td>13</td>
<td>363</td>
<td>493,243</td>
</tr>
<tr>
<td>LITS</td>
<td>Life in Transition Survey</td>
<td>2006–2010</td>
<td>2</td>
<td>35</td>
<td>2</td>
<td>64</td>
<td>67,866</td>
</tr>
<tr>
<td>NBB</td>
<td>New Baltic Barometer</td>
<td>1993–2004</td>
<td>6</td>
<td>3</td>
<td>1</td>
<td>18</td>
<td>21,601</td>
</tr>
<tr>
<td>PA2</td>
<td>Political Action II</td>
<td>1979–1981</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>4,057</td>
</tr>
<tr>
<td>PA8NS</td>
<td>Political Action—An Eight Nation Study</td>
<td>1973–1976</td>
<td>1</td>
<td>8</td>
<td>1</td>
<td>8</td>
<td>12,588</td>
</tr>
<tr>
<td>PPE7N</td>
<td>Political Participation and Equality&lt;sup&gt;e&lt;/sup&gt;</td>
<td>1966–1971</td>
<td>1</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>16,522</td>
</tr>
<tr>
<td>VPCPCE</td>
<td>Values and Political Change&lt;sup&gt;f&lt;/sup&gt;</td>
<td>1993</td>
<td>1</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>4,723</td>
</tr>
<tr>
<td>WVS</td>
<td>World Values Survey</td>
<td>1981–2008</td>
<td>5</td>
<td>89</td>
<td>1&lt;sup&gt;d&lt;/sup&gt;</td>
<td>184</td>
<td>25,682</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1966–2013</strong></td>
<td><strong>89</strong></td>
<td><strong>142</strong></td>
<td><strong>81</strong></td>
<td><strong>1,721</strong></td>
<td><strong>2,289,060</strong></td>
</tr>
</tbody>
</table>

<sup>a</sup>Corresponds to the national surveys.

<sup>b</sup>Full name: Consolidation of Democracy in Central & Eastern Europe.

<sup>c</sup>For Comparative National Elections Project, Eurobarometer, and International Social Survey Programme, only selected survey editions were used.

<sup>d</sup>Common file for the European Values Study and World Values Survey.

<sup>e</sup>Full name: Political Participation and Equality in Seven Nations.

<sup>f</sup>Full name: Values and Political Change in Postcommunist Europe.
Harmonization Process

To create common measures across the pooled survey data, the harmonization team is harmonizing two types of source variables: technical variables, provided by survey administrators, and variables of substantive interest. Currently, the master table contains “new” technical variables for: project ID; Country ID, ISO-1, ISO-2, ISO-3; Wave ID; Year of the study; Country’s administrative unit; Design and post-stratification weights; and Respondent ID within a survey and for all surveys.

So far, our team has created seven substantive target variables, two for the outcome of political protest, and the rest for individual-level characteristics expected to affect it. The theoretical model (see Figure 2) and practical considerations involving the source data informed decisions about which measures to harmonize, and when. The first common measure for protest is a dummy showing whether the respondent participated in demonstrations (yes = 1) or not (no = 0). The other target variable to serve as a dependent variable measures whether respondent signed petition (yes = 1), or not (no = 0). As independent variables, we harmonized information for respondents’ year of birth and age, gender, education, rural/urban, and people’s trust in parliament.

Table 2 shows the number of source variables for a given target variable, and the number of project waves that contain at least one source variable. To illustrate, “participation in demonstrations” appears in 65 out of the 89 project waves; within those 65, most surveys contain only one measure for it; yet in 4 project waves respondents are asked whether they demonstrated regarding specific issues: environmental, human rights, or economic. The situation is similar for signing a petition: in some waves respondents are asked two questions about signing a petition, each referring to a different issue. Our main dependent variables cover more than 1,300 national surveys.

A few comments on independent variables are necessary. Our goal is to have all 2.3 million respondents’ gender and year of birth/age to be able to construct cohorts for historical periods. Although gender was coded differently in various project waves, transforming codes into a unified schema for all national surveys is straightforward. Only 29 waves have a question about respondents’ year of birth. Because this source variable appears in only 33 percent of waves, yet all have age (see Table 2), we use harmonized “age” to construct “year of birth.”

<table>
<thead>
<tr>
<th>Target variables</th>
<th>Maximum number of source variables per wave</th>
<th>Number of waves in which source variables appear</th>
<th>Number of national surveys in which source variables appear</th>
<th>Codes of target variable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation in demonstration</td>
<td>4</td>
<td>65</td>
<td>1,375</td>
<td>yes = 1, no = 0</td>
</tr>
<tr>
<td>Signing petition</td>
<td>2</td>
<td>59</td>
<td>1,321</td>
<td>yes = 1, no = 0</td>
</tr>
<tr>
<td>Gender</td>
<td>1</td>
<td>89</td>
<td>1,720</td>
<td>male = 1, female = 0</td>
</tr>
<tr>
<td>Year of Birth/Age</td>
<td>(1)3</td>
<td>(29)89</td>
<td>(568)1,721</td>
<td>calendar year/years</td>
</tr>
<tr>
<td>Educational categories</td>
<td>18a</td>
<td>76</td>
<td>1,569</td>
<td>codes of ISCED 2011</td>
</tr>
<tr>
<td>Rural/urban</td>
<td>18</td>
<td>86</td>
<td>1,610</td>
<td>rural = 1, urban = 0</td>
</tr>
<tr>
<td>Trust in parliament</td>
<td>2</td>
<td>64</td>
<td>1,248</td>
<td>three scales</td>
</tr>
</tbody>
</table>

*aMost project waves contain up to two source variables dealing with educational categories. In one project wave the only available source variables are country-specific (for 18 national surveys).
The target variable for education is categorical, based on the ISCED 2011 (UNESCO Institute for Statistics 2012) coding schema. ISCED 2011 is an improved version of ISCED 1997, and well-suited for comparing educational levels across different educational systems. In harmonizing education codes we relied mainly on common codes for project waves. In one project wave the only available source variables on education were country-specific codes, giving 18 items for this wave. In addition to constructing the best-suited categories for ISCED 2011 for more than 1,500 surveys, we started harmonizing information on completed years of schooling. The questionnaire item on completed years of schooling is not widely available in the original surveys. However, we intend to deduce years of schooling from educational categories, after relating the latter to current national educational systems.

The harmonization team used survey items asking how much trust or confidence people have in the national parliament to construct the target variable “trust in parliament.” Although the wording of the source questions is similar across surveys, both the length (from 2 to 11 points) and direction (low–high vs. high–low) of the response scales vary. We used different transformation procedures to create three versions of a common measure for trust. The first stretched shorter scales into an 11-point one. The second target variable is a result of squeezing scales into a 0–1 range, where the lowest value of the source question’s scale becomes 0, the highest becomes 1, and the rest is spread evenly between 0 and 1. The third target variable contains information about the relative position of an individual in the distribution of trust in a given project*wave*country (i.e., national survey), showing the share of the national sample that has on average lower trust than that individual. The different versions of the trust target variable are suitable for different types of analyses. For example, the 11-point and the 0 to 1 scales, which are equivalent, facilitate cross-national comparisons of the averages of trust in parliament, whereas the distributional version, which preserves the distribution of original values in national surveys, is better suited for correlational analysis within the national survey.

To deal with differentiation in wording, scales, and/or placement of source variables across surveys, we constructed a range of quality control indicators of the harmonization process, to be included in statistical analyses next to the respective target variable. This solution complements cross-national comparability checks based, for example, on intergroup comparisons of confirmatory factor analyses. We discuss quality controls below, as a key part of the Survey Data Recycling approach.

FORMAL PROCEDURES IN THE DATA HARMONIZATION PROJECT

In light of transparency and replicability considerations, we are developing and documenting a methodology for the entire process of survey data reprocessing. This entails, among others, evaluation of the source data, mapping source and target variables, deciding on specific harmonization rules, and checking the validity and reliability of the harmonized variables. The underlying principle, which we propose under the analytic framework of SDR is that errors and biases in survey data can be dealt with explicitly in substantive analyses via different types of quality control variables (Slomczynski and Tomescu-Dubrow 2015).

In the case of international projects containing surveys conducted on national samples, the core of SDR involves formal procedures around three stages: (1) preparation of the source data for
harmonization, (2) ex post harmonization per se, including evaluation of the quality of harmonization processes, and (3) appending of the pooled survey data with contextual variables from nonsurvey sources. This work requires bringing together three major strands of methodology—data quality, harmonization, and multilevel modeling—which, up to now, have separate scientific literatures. It is complex and still in progress. Detailing it is outside the scope of this article. However, one element of SDR—the relationship of target variable to source variable(s)—is worth discussing because it informs many decisions in the Harmonization Project.

In our approach, the target variable $T$ is considered a function of the source variable(s), $S$, and—in addition—two types of control variables, $Q$ and $H$:

$$T = b_0 + b_1Q + b_2H + e,$$

where $Q$ stands for data quality controls of general survey documentation, specific data description, and data in the computer files; $H$ stands for harmonization quality controls of specific procedures that could influence validity and reliability of $T$; and $e$ denotes random error. If $b_1$ and/or $b_2 > 0$, some intervention is needed to correct for errors and biases in $T$. The solution we propose—which has to be checked empirically—is to partial out (control) the effects of $Q$ and $H$ in statistical analyses.

In the phase of preparing the source data for harmonization, we constructed three types of data quality controls ($Q$). They measure the quality of the general survey documentation (see Schoene and Kołczyńska 2014), the consistency of documentation with data records and the quality of the computer data files (Slomczynski, Powałko, and Krauze 2015).

We should note here that analyses on the harmonization data set uncovered 5,893 nonunique records (i.e., repeated records, “duplicates”) in 162 of the 1,721 national surveys (9.4 percent), distributed unequally across 16 of the 22 international projects (see Slomczynski et al. 2015). Slomczynski et al. defined cases as duplicated “when the set of all answers of a given respondent is identical to that of another respondent” (2015: 1). Overall, they found that national surveys in 80 out of 142 countries contain nonunique records.

We illustrate harmonization quality controls ($H$), created during the stage of ex post harmonization, using the target variable signing petition. While all source questions unequivocally ask about the activity of signing a petition, the time frame of reference for carrying it out varies across project waves. For example, in some waves respondents were asked “Did you sign a petition in the last six months?” while in others the question was “Have you signed a petition in the last seven years?” and in still others “Have you ever signed a petition?” Hence, we constructed a control variable for the time frame of the original questions about signing petition. Its values are $1$—one year or less, 3, 5, 7, 10, and 11—years, and $12$—ever.

We maintain that researchers should include this indicator, next to the target variable, in substantive analyses. This solution does not preclude, but rather accompanies, cross-national comparability checks based, for example, on intercountry and intergroup comparisons. For example, in some countries signing a petition can be rare since it is substituted by different forms of protest. This seems to be especially relevant for analyzing countries with weak petition-signing culture, as is the case in new democracies (see Inglehart and Catterberg 2002). Similarly, in the case of intergroup comparisons, one should remember that signing a petition is not in the regular repertoire of protest behavior of the segment of society where illiteracy prevails. Analyzing intercountry and intergroup comparability of indicators complements the harmonization process.
The relevance of both survey quality \((Q)\) and harmonization \((H)\) controls for substantive results needs to be tested empirically, which is intrinsic to the SDR research agenda. Because quality indicators could be defined on different levels, such as the national survey or the project-wave, the general Equation (2) linking the target variable to the source variable(s) must include subscripts reflecting the complex structure of the data. This structure calls for the use of multilevel modeling techniques.

CONCLUSIONS

The Harmonization Project is ongoing and thus a work in progress. To test the theoretical model of political protest we combined individual-level information from international projects with country-level information from nonsurvey sources into Big Data. The master table of 2.3 million respondents from a total of 142 countries or territories and spanning the period 1966 to 2013, and its combination with macro-level variables, fits a popular definition of Big Data as an unusually large number of units of observation and a large number of variables from a diversity of sources (Mayer-Schönberger and Cukier 2013). However, Big Data are often understood as information that is “out there,” produced independently of researchers who “just gather” the data and use them in an exploratory manner. In contrast, survey data are produced by researchers, and, combined with other data, are well-suited for applying sociology/political science theory to test specific hypotheses. We use the term “big data” sensu largo: basic units of observation (from different projects) in millions, and variables of interest (from different sources) in hundreds.

We agree with Markus Hadler and his colleagues (2015) who point out that, on the one hand, researchers are confronted with large sets of data offering many different options in terms of geographical coverage, survey topics, and time points but, on the other, the basic methodological problems of functional equivalence, selection of countries, and the quality of data remain. Before we can analyze our Big Data in an integrated format, various methodological problems inherent in cross-national surveys and in ex post harmonization have to be resolved. Among them, dealing with the methodological variability in original materials and ensuring quality of the target variables rank high.

We are in the process of formalizing the solutions to these complex challenges within the SDR framework. A key aspect involves constructing quality-control indicators, which become part of the integrated data set. According to the SDR logic, control variables for the quality of the source data provide means for reprocessing existing cross-national projects in ways that minimize the “messiness” of data built into total survey error, while harmonization controls, which record (and reveal) the decisions of ex post harmonization, facilitate validity and reliability assessments of the target variables.

ACKNOWLEDGMENTS

We acknowledge comments on the project from Edward Crenshaw, Tadeusz Krauze, Dean Lillard, Malgorzata Mikucka, Francesco Sarracino, and Sandy Marquart-Pyatt.
FUNDING

This project was conducted by the Institute of Philosophy and Sociology of the Polish Academy of Sciences (IFiS PAN) and the Mershon Center for International Security Studies at Ohio State University (MC OSU), with funding from the National Science Centre, Poland (grant number 2012/06/M/HS6/00322). The Cross-National Studies: Interdisciplinary Research and Training program (CONSIRT consirt.osu.edu) hosts the project. The project team includes Kazimierz M. Slomczynski (PI), J. Craig Jenkins (co-PI), Irina Tomesu-Dubrow (co-PI), Joshua K. Dubrow, Marta Kołczyńska, Matt Schoene, Ilona Wysmulek, Olena Oleksiyenko, Przemek Powałko, and Marcin Zielinski. In addition, Anastas Vangeli, Jakub Wysmulek, and Anna Franczak helped in technical matters. Zbigniew Sawiński served as consultant for harmonizing some variables discussed in this article.

NOTES

1. We assume that most of these variables could be operationalized in different ways. We use notations similar to those used in popular HLM texts such as Hox (2010), Gellman and Hill (2007), and Raudenbush and Bryk (2002), and adjust the explanation of coefficients to cross-national research where individuals are nested in countries.

2. We refer to the selected projects as well-known on the basis of publication records and their effect on the social-science disciplines. Data gathered in March 2015 show that the number of publications from projects listed in Table 1 was 2,087 at the Web of Science; 11,746 at the projects’ home pages; and from 20,051 to 116,377 (depending on how strict limitations of the searches were) at Google Scholar. At the same time, the number of citations of the projects in Table 1 was 19,726 (Web of Science).

3. For practical reasons, we stopped adding new data in the second quarter of 2014.

4. Because of the thematic coverage criterion, we include only survey waves that contain relevant questions on protest behavior and/or democratic values; hence, not all waves of ISSP, Eurobarometer, and Comparative National Elections Project are in our data.

5. Some of the international survey projects require agreeing to the “click license” stipulating “not to pass along the data to third parties.” For these projects we are seeking permission to include technical and source variables in the master file that are necessary for a replication of our work.

6. Many of the national surveys included in our project are prior to 2011 and employ ISCED 1997. Although the 1997 and 2011 ISCED coding differs, the conversion process is possible with some additional information, generally provided in country-specific classifications.

7. The assumption is that particular values of shorter scales correspond to a range of values on the 0–10 scale. For an $n$-point scale with $k$ values ranging from 1 to $n$, $k$ was recoded to $10/(n \times 2) + (k - 1) \times 10/n$.

8. For an $n$ point scale, for values $k$ ranging from 1 to $n$, $k$ was recoded to $(k - 1) + \frac{1}{n+1}$.

9. For an $n$-point scale with $k$ values ranging from 1 to $n$, where $X_k$ is the distribution of the variable, $k$ was recoded to $\sum_{i=1}^{k-1}X_i + \frac{k}{2}$. The correlation between the two versions of the trust target variable (11-point scale vs. distribution-based variable) is 0.881 for the pooled 1,721 surveys. Within survey, the correlation ranges from 0.914 to 0.99988; the mean correlation is 0.985.

REFERENCES


