

*Harmonization:*  
Newsletter on Survey Data  
Harmonization in the Social Sciences

Editors  
Irina Tomescu-Dubrow  
and  
Joshua Kjerulf Dubrow  
CONSIRT

[consirt.osu.edu/newsletter/](http://consirt.osu.edu/newsletter/)  
ISSN 2392-0858

## One Step at a Time

The *Harmonization* newsletter presents the latest news from the community of scholars, institutions, and government agencies who work on harmonizing social survey data and on other projects focusing on the methodology for comparative and cross-national studies. One step at a time, this community moves forward the interdisciplinary field of survey data harmonization in the social sciences.

In this issue, we feature a variety of news on big data, harmonization, and data quality. **The Harmonization Project** published its data on Harvard's Dataverse, and The Ohio State University opened the **Translational Data Analytics Institute** focused on big data. In the articles this newsletter features, **Koen Beullens and colleagues** summarize their European Social Survey data quality report, **Verena Ortmanns and Silke Schneider** present their latest research on cross-national harmonization of educational attainment variables, **Kea Tijdens** discusses measurement of occupations in multi-country surveys, **Irina Tomescu-Dubrow** and colleagues discuss metadata on survey quality, and **Anna Turner** argues for more and better sociological use of Google search data.

The harmonization community continues to present their research at conferences and workshops around the world. In this issue, we have an announcement of a SERISS workshop in Amsterdam, a report from a presentation at CSDI in Mannheim, and a summary of a conference organized by CONSIRT at The Ohio State University's Mershon Center on the methodology of studying protest.

As always, we invite all scholars interested in survey data harmonization to read our newsletter and contribute their articles and news to future editions.

## In This Issue

[News, p. 2](#)

[Articles, p. 3](#)

[Quality Assessment in the European Social Survey, p. 3](#)

[Quality and Comparability of Education Data: New Publications, p. 5](#)

[Harmonization of the Measurement of Occupations in Multi-country Surveys, p. 6](#)

[Quality Controls of Data from International Survey Projects, p. 9](#)

[Surveillance, Privacy, and Leaks: Google Big Data in Cross National Perspective, p. 14](#)

[Conferences and Workshops, p. 16](#)

[Contact Us, p. 19](#)

**Acknowledgements**

**Editors thank Marta Kołczyńska for technical assistance**

# News

---

## The Harmonized Dataset Published in Harvard University's Dataverse

by Survey Data Harmonization Team, Polish Academy of Sciences and The Ohio State University

The Harmonization Project from OSU and the Polish Academy of Sciences ([dataharmonization.org](http://dataharmonization.org)) has published their harmonized dataset on Harvard University's Dataverse, where it can be downloaded from free of charge.<sup>1</sup>

The SDR Master Box consists of five data files - containing harmonized technical and substantive variables, metadata in the form of quality control variables measured at different levels of the source data, metadata pertaining to the harmonization process, and time-varying macro-level characteristics of respondents' countries. Corresponding documentation accompanies each of the following files:

- (1) master file (MASTER) with individual-level data from cross-national surveys,
- (2) country-level file (PLUG-COUNTRY),
- (3) country-year-level file (PLUG-COUNTRY-YEAR),
- (4) survey-level file (PLUG-SURVEY), and
- (5) wave-level file (PLUG-WAVE).

The MASTER file is the core of the Master Box and contains harmonized target variables, harmonization metadata, as well as individual-level control indicators for the quality of the source data. The other PLUG files contain contextual data and control variables for the quality of the source data measured at the survey and the project-wave levels, respectively.

## Solving Complex Challenges with Data Analytics at Ohio State

by Translational Data Analytics Institute, The Ohio State University

Solving today's complex challenges requires data analytics and public-private collaborations like never before. Through an investment of more than \$125 million in a new Translational Data Analytics Institute (TDAI), The Ohio State University has committed to advancing translational data analytics education, research, and scholarship to utilize Big Data for Good.

---

<sup>1</sup> Slomczynski, Kazimierz M.; Jenkins, J. Craig; Tomescu-Dubrow, Irina; Kolczyńska, Marta; Wysmulek, Ilona; Oleksiyenko, Olena; Powalko, Przemek; Zieliński, Marcin W., 2017, "SDR Master Box", [doi:10.7910/DVN/VWGF5Q](https://doi.org/10.7910/DVN/VWGF5Q), Harvard Dataverse, V1.

The use of the term “translational” reflects a fundamental shift toward utilizing data science and analytics in solving issues of global importance – translating research into solutions. It is based on the principle that the best solutions are those that draw upon the collective expertise of people across disciplines. Solutions such as using satellite and drone data to develop faster recovery responses after natural disasters, using statistical techniques to protect the cassava plant in east Africa, and exploring what factors most influence children’s health when giving vaccines.

*“Translational” reflects a fundamental shift toward utilizing data science and analytics in solving issues of global importance.*

TDAI works every day to enable collaboration and innovation in translational data analytics, co-develop externally responsive translational data analytics solutions with industry and community partners, expand the workforce capable of delivering translational data analytics solutions, and build a sustainable core of data science and analytics scholarship. We are helping to establish Ohio as a global hub for the study, development, and application of translational data analytics solutions with:

- A virtual Data Commons currently in development that will connect users for sharing data sets, knowledge, and software
- A newly renovated physical hub in the heart of campus opening spring 2018
- Forums, speakers, seminars, challenges, technology demos and workshops, and national consortia
- Faculty appointments, seed grants, and support for student traineeships
- Curriculum planning and integration
- A portfolio of projects that address faculty and partner needs

Read more about TDAI is using Big Data for Good at <https://discovery.osu.edu/tda/>.

## Articles

---

### Quality Assessment in the European Social Survey

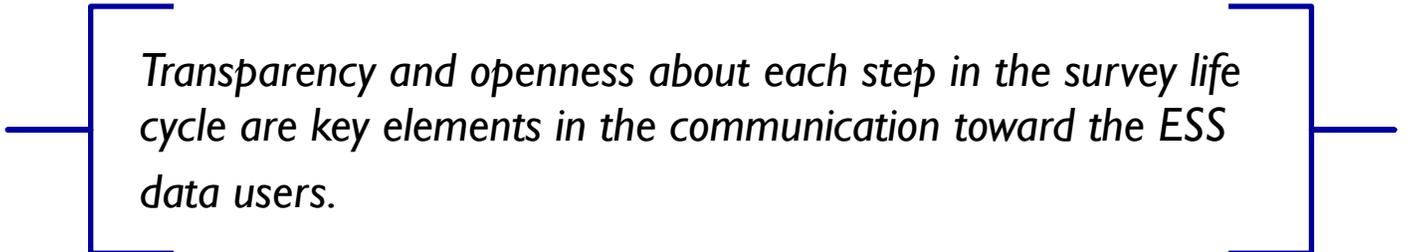
by Koen Beullens, Geert Loosveldt, and Caroline Vandenplas,  
Centre for Sociological Research – University of Leuven, Belgium

The European Social Survey (ESS) is a biennial cross-national survey that covers a wide range of topics such as well-being, attitudes towards immigration, politics, etc. Methodologically, the ESS has set high quality standards regarding sampling, (non)response, translation, questionnaire design,

interviewer management and other issues related to the harmonization principles of cross-national surveying. Since round 6 in 2014, the Central Scientific Team (CST) of the ESS has therefore taken the initiative to assess the quality of the survey, overall and on the country level.

In this context, transparency and openness about each step in the survey life cycle are key elements in the communication toward the ESS data users. All the design characteristics of the survey are clearly documented: how the sampling has been done in all countries, how the interviewers have been paid, the number and nature of questionnaire pretests, etc. Even more, the entire fieldwork history per sampled case is made available by means of the contact form dataset. This information can easily be found on [europeansocialsurvey.org](http://europeansocialsurvey.org). This openness and transparency principle enables to a reasonable extent the assessment of the quality of the survey process and its outcome.

Inspired by the principles of Total Quality management (TQM) and Total Survey Error (TSE), the quality assessment focuses on process as well as output aspects of survey quality. Typical process-related issues concern the different elements in the survey life cycle such as the sampling strategy, the translation process, interviewer training, interviewer capacity, monitoring fieldwork activities, timeliness, pretesting, efforts to reduce nonresponse error or data cleaning. The outcome aspects relate to response rates, indications of nonresponse bias, interviewer effects or other measurement issues such as straightlining or item-nonresponse.



*Transparency and openness about each step in the survey life cycle are key elements in the communication toward the ESS data users.*

After each ESS round, the National Coordinators receive a country specific report that summarizes their performance on these process and outcome aspects for their country and suggests relevant points of interest where the country might improve itself in the next ESS round. The reports are also suggestive regarding quality issues that relate to all countries and may result in proposals for general quality improvement. In this light, particular initiatives have been taken to standardize interviewer training (since substantive interviewer effects have been found in many countries). Future initiative will also concentrate on backcheck procedures, interviewer remuneration and nonresponse bias.

*Koen Beullens, Geert Loosveldt, and Caroline Vandenplas and researchers at the Centre for Sociological Research, University of Leuven, Belgium. The quality report for the 7<sup>th</sup> round of ESS can be found at [europeansocialsurvey.org](http://europeansocialsurvey.org). Note that the report only contains the general quality assessment; country specific reports have only been made available to National Coordinators.*

## Data Quality and Cross-national Comparability of Educational Attainment: New Publications

by Verena Ortmanns and Silke Schneider, GESIS

During recent decades, cross-national comparative research has grown tremendously, both in quantity and quality. Through the increased availability of various types of cross-national survey data, many research questions can today be tackled from a comparative point of view. The credibility of comparative studies, however, hinges on the cross-national comparability of the data and their consistencies.

The educational attainment variable, a socio-demographic variable covered by virtually every survey of individuals, can be used to assess data quality of cross-national surveys. In 2016, Ortmanns and Schneider published two papers comparing the data quality of the education variable within and between multiple cross-national surveys.

*The credibility of comparative studies hinges on the cross-national comparability of the data and their consistencies.*

Previous research has shown that substantial inconsistencies in the distribution of the education variable coded in the International Standard Classification of Education (ISCED) exist across time and surveys, threatening cross-national and cross-survey comparability. Ortmanns and Schneider updated this work by using recent data and by increasing number and scope of surveys. In both papers, they use the Duncan's Dissimilarity Index to compare the distributions of the harmonized education variable with the respective education distributions of high-quality benchmark data in country-year combinations. Their research entitled "Harmonization still failing? Inconsistency of education variables in cross-national public opinion surveys" (Ortmanns and Schneider 2016a) analyzes data from four cross-national, public opinion surveys including the European Social Survey (ESS), the Eurobarometer, the International Social Survey Programme (ISSP), and the European Values Study (EVS). The education distributions are compared over time and across surveys using the ESS as benchmark data.

In the article "Can we assess representativeness of cross-national surveys using the education variable?" (Ortmanns and Schneider 2016b) data from the European Labour Force Survey (EU-LFS), the European Union Statistics on Income and Living Conditions (EU-SILC), and the Programme for the International Assessment of Adult Competencies (PIAAC) are added to previous analysis. Here data of the EU-LFS are used as benchmark. This article includes a second, more qualitative, analysis to identify and to explain the observed inconsistencies by looking at both measurement errors in the education variables and errors of representation. A lot of examples are provided to illustrate errors related to measurement instruments, data processing, and non-observation. Both papers close with a number of suggestions on how to improve the consistency of educational attainment variables.

For more information, see:

Ortmanns, V., & Schneider, S.L. (2016a). Harmonization still failing? Inconsistency of education variables in cross-national public opinion surveys. *International Journal of Public Opinion Research*, 28(4), 562-582. doi: 10.1093/ijpor/edv025

Ortmanns, V., & Schneider, S.L. (2016b). Can we assess representativeness of cross-national surveys using the education variable? *Survey Research Methods*, 10(3), 189-210. doi:10.18148/srm/2016.v10i3.6608

*Verena Ortmanns is a Ph.D. student and junior researcher at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany.*

*Silke L. Schneider is a senior researcher and consultant at GESIS – Leibniz Institute for the Social Sciences in Mannheim, Germany, and she is the head of the CAMCES-project. She focuses on research and consulting on the measurement of socio-demographic variables in survey research.*

## Harmonization of the Measurement of Occupations in Multi-country Surveys

by Kea Tijdens, University of Amsterdam, Netherlands

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) is a four-year project that aims to strengthen and harmonise social science research across Europe (EU-Horizon 2020, no 654221, 2015-19). Work Package 8 of SERISS aims to provide cross-country harmonized, fast, high-quality and cost-effective coding of open ended questions on respondents' occupations, industries and education into international standardized classification systems.

This article solely addresses the measurement and coding of occupations. It builds on work done for a previous project, InGRID, specifically its paper "Develop methods to facilitate the EU-wide measurement of occupations in web surveys" (Tijdens, 2015).

***SERISS develops a cross-country harmonised, fast, high-quality  
and cost-effective coding module for occupations.***

The measurement of occupations in a cross-cultural and cross-national context is sometimes cumbersome, not sufficiently standardized and often expensive because of office coding. Building on the current technology and the partners' experiences, SERISS Work Package 8 develops a cross-country harmonised, fast, high-quality and cost-effective coding module for occupations. The module uses a large multi-lingual database with thousands of occupations, all coded according to the international ISCO-08 classification. The module will firstly facilitate multi-country surveys such as ESS, GGP and SHARE, but ultimately aims for reaching out to a global audience. The responsible partner is the University of Amsterdam (UvA).

The multi-lingual database of occupations has a long history. Starting in 2000 as a web survey

in the Netherlands posted on a frequently visited website with information on wages and working conditions, a database of occupations was gradually developed for use as a look-up table for the survey question ‘What is your occupation?’ This database, called WISCO database, grew because occupational titles were added. With more occupations in the look-up table, a search tree was developed to facilitate respondents to search the database for their occupational title. From 2004 onwards, the WageIndicator web survey expanded to neighbouring countries due to an EU funded FP6 project. Partners added translations of titles, as well as occupational titles not yet present. More countries joined the WageIndicator, and in these cases new translations and new titles were added.

In 2009, the Database of Occupations was revised and all occupational titles were coded according to ISCO-08 classification. The database was uploaded in the WageIndicator web survey for all countries, using a three-level search tree. The database and its search tree were also implemented in the Salary Checks. The database included 1,594 distinct occupational titles with translations for approximately 30 languages for almost 50 countries in and outside Europe with a WageIndicator website. With millions of web visitors and over a hundred thousand respondents per year, the occupation database and its search tree were tested extensively. Web visitors and survey respondents send relatively few emails to WageIndicator about the occupations, pointing to a satisfactory search tree and its related list of occupations.

In the years after 2009, occupations have been added to the database, mostly because web visitors requested so in their emails. For the UK, a number of management occupations and for the Czech Republic and Slovakia medical specialists have been added and for Germany the skill levels for some skilled occupations have been further detailed by distinguishing occupations at university and higher vocational level. By 2015 the database held 1,896 occupational titles, of which 132 were country-specific. Apart from the latter 132, almost all titles were translated in the 43 languages of the almost 90 national WageIndicator websites and web surveys. The number of occupational titles varied slightly across countries, because in some countries some occupation titles could not be translated or distinct occupational titles in the source list were translated similarly. Figure 1 shows a cut out of the database.

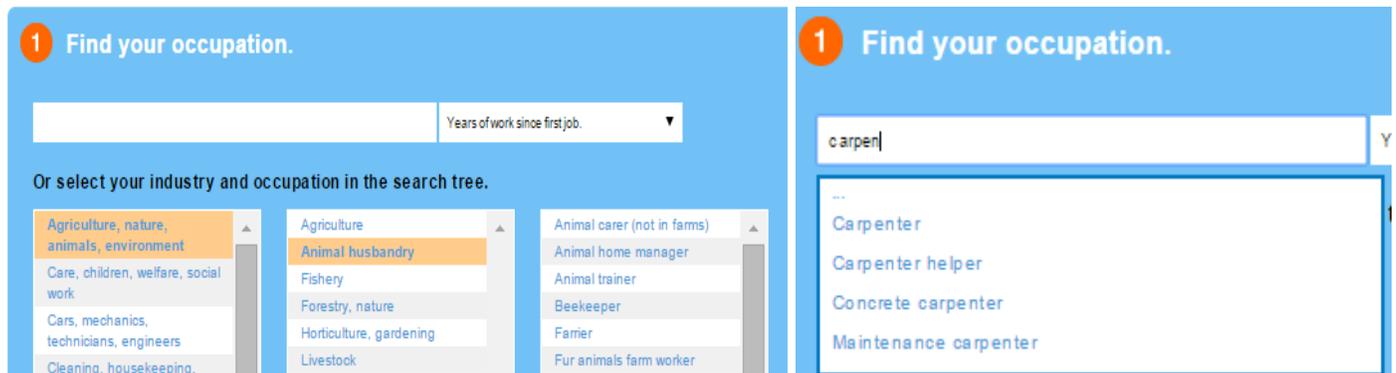
Figure 1. Screenshot of the 2015 WISCO database of occupations.

1	ISCO0813	Master label	am_ET	ar	az_AZ	ba_ID
2	110010000000	Commissioned officer armed forces	የጦር ኃይሎች ባለሙያዎች መኮንን	ضباط في القوات المسلحة	Hərbi qüvvələrin zabiti	Perwira angkatan darat
3	110020000000	Military operations leader	የጦር ኃይሎች ዘመቻዎች መሪ	قائد العمليات العسكرية	Hərbi əməliyyatların m	Pimpinan operasi militer
4	111101000000	Member of parliament, legislator	የፓርላማ አባል፣ የሕግ አውጪ	عضو في البرلمان ، مشرع	Parlament üzvü	Pembuat undang-undang
5	111102000000	Mayor, alderman		العمدة ، رئيس مجلس بلدي	Mer, icra hakimiyyətinin	Mayor, anggota dewan kota
6	111113000000	City councillor, county councillor	የከተማ ምክር ቤት አባል፣ የገር ምክር ቤት	عضو مجلس محلي أو القليبي/بلدية	Şəhər şura üzvü, Region ş	Anggota dewan kota
7	111202000000	Public administration manager	የሕዝብ አስተዳደር ማኔጀር	مدير الدائرة العامة	Administrasiya rəisi, dövlət	Manajer administrasi publik
8	111203000000	Senior non-legislative official	ከፍተኛ ሕግ አውጪ ያልሆነ ባለሥልጣን	كبار المسؤولين الحكوميين من غير	Qeyri-dövlət müəssisəsini	Pejabat senior bukan pembuat
9	111204000000	Senior government official	ከፍተኛ የመንግስት ባለሥልጣን	مسؤول حكومي كبير	Dövlət müəssisəsini işçisi	Pejabat senior pemerintahan
10	111206000000	Diplomat	ዲፕሎማት	دبلوماسي	Diplomat	Diplomat
11	111301000000	Traditional chief or head of village	ባሕላዊ አለቃ ወይም የመንግሥት ኃላፊ	العمدة أو رئيس القرية	Kənd sovetinin nümayənd	Kepala kampung/suku
12	111401000000	Employers' organisation official	የግሪዎች ድርጅት መኮንን	مسؤول منظمة أرباب العمل	Məşğulluq xidməti nümay	Pejabat senior organisasi peng
13	111402000000	Humanitarian organisation official	የሰብአዊት ድርጅት መኮንን	مسؤول منظمة إنسانية	Humanitar yardım təşkilat	Pejabat senior organisasi kema
14	111403000000	Political party official	የፖለቲካ ፓርቲ ባለሥልጣን	مسؤول /الأحزاب السياسية	Siyasi partiya nümayəndə	Pejabat senior partai politik

Note: The first column shows the ISCO-08 code (first four digits are the ISCO code); the second column shows the source label, the remaining columns show the translations for Amharic, Arabic, Azeri and Indonesian.

In 2012, the search tree in the web survey and the Salary Check was extended with a semantic matching tool (Figure 2). Semantic matching allows visitors to self-identify their occupation by typing text whereby matches with words in the list of occupations for their particular language are instantly shown. Visitors can then select the most relevant match.

Figure 2. Screenshot for the tool to self-identify occupation with a search tree and with semantic matching.



Source: *WageIndicator Salary Check for Great Britain* [wageindicator.co.uk](http://wageindicator.co.uk), accessed 15-12-2015

Although more than 1,700 occupational titles may seem a large number, one has to take into account that a labour market in any country can easily include 10,000 or more job titles. The use of the database for self-identification in web surveys would therefore profit from extending the number of occupations. In 2015, the SERISS project provided funding to extend the database to almost 5,000 titles and to 99 countries, and to develop an Application programming interface (API) for the measurement of occupations in web surveys. The choice of countries and languages included arguments related to EU-28 and associate countries, to OECD and associate countries, to countries with the five most spoken languages outside these areas: Russian, Mandarin, Arabic, Hindi and Bahasa, and to countries where the SERISS infrastructures with their multi-country surveys had plans for new waves. The database services the measurement of occupations in 99 countries with 47 languages.

Survey holders can use the `occupation_API` for their surveys. This use is for free until the end of the SERISS project, but it remains available after the project. Details are available at the website [surveycodings.org](http://surveycodings.org), accessible from Autumn 2017 on. For more information, see also the Work Package 8 of SERISS, as well as previous work on the database, listed in the references.

*Kea Tijdens is Professor of Women's Work at the Department of Sociology, Erasmus University Rotterdam and a Research Coordinator at ALAS/University of Amsterdam.*

## References

- Belloni M, Brugiavini A, Meschi E, Tijdens KG (2016) Measurement error in occupational coding: an analysis on SHARE data, *Journal of Official Statistics*, Volume 32, Issue 4, Pages 917–945
- Tijdens KG (2014) Drop-out rates during completion of an occupation search tree in web-surveys, *Journal of Official Statistics*, 30 (1), pp. 23–43

Tijdens KG (2015) Self-identification of occupation in web surveys: requirements for search trees and look-up tables, *Survey Methods: Insights from the Field*, <http://surveyinsights.org/?p=6967>

Tijdens KG (2015) *The design of a tool for the measurement of occupations in web surveys using a global index of occupations*. Leuven, Working paper InGRID project, inclusivegrowth.be

## Quality Controls and Their Application to Substantive Analyses of Data from International Survey Projects

by Irina Tomescu-Dubrow, Kazimierz M. Slomczynski, and Marta Kołczyńska,  
Polish Academy of Sciences, CONSIRT, and The Ohio State University

This article proposes ways to account for unequal quality of source data, which scholars are bound to encounter in the process of ex-post harmonization of cross-national surveys. It is work in progress that originated in the Harmonization Project ([dataharmonization.org](http://dataharmonization.org)), where we put together a multi-country multi-year dataset suitable for studying individual and contextual determinants of peoples' propensity to engage in political protest. Our team pooled information from 22 well-known international projects into a dataset of 2,289,060 respondents from 1,721 national surveys conducted between 1966 and 2013 in a total of 142 countries and territories. The harmonized SDR dataset is available via Dataverse (see also first item in this newsletter).<sup>2</sup>

To systematically evaluate the quality of the 1721 surveys, we apply rules developed within the Survey Data Recycling analytic framework. The main idea in SDR is to construct control variables dealing with different aspects of the quality of the source surveys, and separate controls for transformations of source variable as part of harmonization. This way researchers can deal with methodological biases and errors in survey data explicitly. In this context, methodological biases and errors in survey data are understood as consequences of deviations from standards of documenting and preparing survey data suggested in the specialized literature (e.g. Biemer and Lyberg 2003), and inter-survey differences in harmonized items.

***We construct metadata that measure variability in the quality of the source survey data.***

On the basis of Total Survey Error (TSE), Total Survey Quality (TSQ) and Total Quality Management (TQM) frameworks, SDR identifies three relevant dimensions of survey quality: (a) the quality of surveys as reflected in the general survey documentation, as lack or inadequate information in the general survey documentation reduces confidence in the data ; (b) the degree of

<sup>2</sup> Slomczynski, Kazimierz M.; Jenkins, J. Craig; Tomescu-Dubrow, Irina; Kolczyńska, Marta; Wysmulek, Ilona; Oleksiyenko, Olena; Powalko, Przemek; Zieliński, Marcin W., 2017, "SDR Master Box", [doi:10.7910/DVN/VWGF5Q](https://doi.org/10.7910/DVN/VWGF5Q), Harvard Dataverse, V1.

consistency between the description of the data with the data records in the computer files, since inconsistencies between the two, as a specific type of processing errors, can affect the overall usability of the survey; and (c) the quality of the data records in the computer files, since errors in data files lead to possible distortion of empirical results based on those data.

For each dimension, we construct indicators that measure the methodological variability in the source survey data, as follows: (1) variables that describe the quality of data documentation dealing with sampling design, preparing the questionnaire, pre-testing and fieldwork control (Schoene and Kolczyńska 2014; Kolczyńska and Schoene *forthcoming*); (2) variables that capture inconsistencies between survey documentation and records in the computer data files (Oleksiyyenko, Wysmulek, and Vangeli, *forthcoming*); and (3) variables describing errors or biases in data records in the computer file such as frequencies of erroneous respondents' IDs, non-unique records (duplicates), missing data across socio-demographics, and erroneous weights (Zieliński, Powalko, and Kolczyńska, *forthcoming*). Table 1 provides an overview of the three sets of metadata for source survey quality, which are available in the SDR dataset (see Footnote 1).<sup>3</sup>

Analyzing the documentation for the 1,721 surveys in the Harmonization Project shows that 24.5% surveys involve features of quota sampling, and almost an equal number (24.4%) were conducted on the basis of sampling schemes that involved a random route component.<sup>4</sup> Simple or stratified random samples were reportedly used in 5.7%, while multistage individual register and multistage address register were used in 7.4% and 10.2% of the surveys, respectively. The rest contains three categories: multistage unspecified, no sufficient information or missing information.

Table 1 includes summary indicators that we discuss in the current analysis. In the case of the quality of the data as reflected in the documentation (Panel I, items (a)-(e)) and computer records (Panel III, items (a)-(d)) we use indexes that are sums of binary variables. The index measuring (in)consistency of data description with computer records of survey data was calculate as the average number of processing error per variable, since not all surveys contain all variables that were examined (gender, age, year of birth, education level, years of schooling, trust in parliament, and participation in demonstrations; see SDR PLUG-WAVE documentation file on Dataverse).<sup>5</sup> For all indexes we use in this paper, the higher their value the higher the quality of the survey.

Table 2 presents the distribution of the survey data quality indexes across the 1,721 surveys in the SDR dataset, and the extent to which they relate to each other. The values of the pairwise correlation coefficients between the different indexes of survey data quality – pertaining to documentation (D), consistency of the data and the documentation (I), and computer records of the data (R) are low, supporting the claim that the indexes measure distinct dimensions of survey quality.

---

<sup>3</sup> Some of the dichotomous quality control variables in Table 1 could be expressed on the interval or ratio scales, e.g., how many checks of the questionnaire adjustment were made or the percent of unit-nonresponses.

<sup>4</sup> The description of the coding of sampling methods is available in Slomczynski, Powalko, and Krauze 2017 (pp. 6, 8). In analyzing the documentation, we found instances when features of quota sampling were reported together with features of different methods, random route especially. We coded such instances as quota.

Table 1. Measures of survey quality.

<b><u>I. General Survey Documentation: How were the data collected?</u></b>	
Type of the sample	0 – No information 1 – No sufficient information 2 – Quota 3 – Random route 4 – Multistage unspecified 5 – Multistage address register 6 – Multistage individual register 7 – Simple/stratified random
(a) Does the survey documentation describe accurately the sampling procedure?	0 = No 1 = Yes
(b) Does the survey documentation provide information on the response rate?	0 = No 1 = Yes
(c) Was the quality of questionnaire translation checked in some systematic way (e.g., back-translation, TRAPD)	0 = No 1 = Yes
(d) Is there any evidence that the instrument (questionnaire) was pre-tested?	0 = No 1 = Yes
(e) Is there evidence of fieldwork control?	0 = No 1 = Yes
Summary index of items (a) – (e)	From 0 to 5
<b><u>II. Specific Data Description: How were the data defined?</u></b>	
(a) Is the description of the data consistently provided in English?	No = 0 Yes = 1
(b) Are all variable values within the legitimate range?	No = 0 Yes = 1
(c) Are variable labels the same in the codebook and in the data file?	No = 0 Yes = 1
(d) Are value labels the same in the codebook and in the data file?	No = 0 Yes = 1
(e) Are different types of missing values explained?	No = 0 Yes = 1
Summary measure of items (a) – (e) standardized for number of questions/variables included in the analysis	From 0 to 1.33
<b><u>III. Computer Data File: Were the data formally correct?</u></b>	
(a) Are the data free of non-unique records?	No = 0 Yes = 1
(b) Are design/post-stratification weights free of formal errors?	No = 0 Yes = 1
(c) Is the proportion of missing values for gender and age below 5%?	No = 0 Yes = 1
(d) Do survey cases (respondents) have unique identification numbers (IDs)?	No = 0 Yes = 1
Summary index of items (a)-(d)	From 0 to 4

<sup>5</sup> [dataverse.harvard.edu/file.xhtml?fileId=3006263&version=RELEASED&version=.0](http://dataverse.harvard.edu/file.xhtml?fileId=3006263&version=RELEASED&version=.0)

Research on fabrication and falsification of survey data (Bredl, Storfinger, and Menold 2011; de Hass and Winker 2014) and fake and fraudulent interviews (Schafer, Schrapler, and Muller 2004; Blasius and Thiessen 2012) provides useful insights with regard to the relevance of survey quality measures for empirical studies. Evidence referred in these papers indicates that “bad cases” usually lead to overreporting of socially desirable opinions and behaviors (and hence upwardly biased sample averages or proportions), and overestimation of expected correlations.

Table 2. Distributions of the survey data quality indexes in the SDR dataset.

Survey data quality measures	Mean	SD	% of national surveys in highest category
D = General Survey Documentation, summary index, 0-5 (high=5)	2.53	1.68	20.8
I = Specific Data Description, index 0-1.33 (high=1.33)	1.16	0.26	48.7
R = Data Records in Computer File, summary index, 0-4 (high=4)	3.36	0.67	46.5
$r_{DI} = 0.160^*$ ; $r_{DR} = 0.234^*$ ; $r_{IR} = -0.018$ ; $n = 1,721$ national surveys			

In Table 3 we provide initial results pertaining to the first hypothesis that survey quality is not randomly distributed with respect to the substantive target variable participation in demonstrations. We regress participation in demonstrations on each source data quality index, in two-level regression models. In column 2 the dependent variable is the proportion of people who declared having demonstrated in a given country-year, and countries form the 2<sup>nd</sup> level grouping variable. In column 3, the dependent variable is respondents’ odds of participating in demonstrations, and the 2<sup>nd</sup> level grouping variable is country-years.

Table 3. Two-level regression models of participation in demonstrations in the last 10 years or ever<sup>6</sup> on the indexes of source data quality.

Data quality index	Level 2: countries (119)	Level 2: country-years (704)
	b for proportions	b for logit
Data documentation (index, 0-5)	0.002	0.019
Accuracy of data description (index, 0-1.33)	0.044**	0.533**
Quality of computer records (index, 0-4)	-0.026**	-0.208**

We find that two of the three source quality metadata – accuracy and quality of computer records – are not randomly distributed with respect to participation in demonstrations. This represents a first

<sup>6</sup> The wording of the question about past participation in demonstrations varies across survey projects, most notably in the time mentioned in the question text, which ranges from „last 12 months” to „ever”, which results in responses that are not readily comparable. In this analysis we selected surveys that ask whether the respondent has participated in a demonstration “ever”, which is also the most common formulation of this item in the survey projects we have analyzed.

indication that survey quality variables matter in substantive research. As a next step, we will check whether the data support the second, ‘strong’ expectation, that low quality of the source data influences the association between participation in demonstrations and respondents’ education.

*Irina Tomescu-Dubrow is Associate Professor at the Institute of Philosophy and Sociology, Polish Academy of Sciences (PAN). She is Program Manager at Cross-National Studies: Interdisciplinary Research and Teaching Program (CONSIRT), of OSU and PAN.*

*Kazimierz M. Slomczynski is Professor Dr. Hab. at IFiS PAN and Emeritus Professor, Department of Sociology, The Ohio State University (OSU). He directs the Cross-National Studies: Interdisciplinary Research and Training program (CONSIRT) of OSU and PAN.*

*Marta Kolczyńska is a Ph.D. candidate at the Department of Sociology, The Ohio State University, affiliated with the Cross-National Studies: Interdisciplinary Research and Training program (CONSIRT) of OSU and PAN.*

## References

- Biemer, P. P. and Lyberg, L. E. (2003) *Introduction to survey quality*. New York: John Wiley & Sons, Inc.
- Blasius, J. and Thiessen, V. (2012) *Assessing the quality of survey data*. London: SAGE.
- Bredl, S., N. Storfinger and Menold, N (2011) [A literature review of methods to detect fabricated survey data](#), No 56, Discussion Papers, Justus Liebig University Giessen, Center for international Development and Environmental Research (ZEU).
- de Haas, S. and Winker, P. (2014) Identification of partial falsifications in survey data. *Statistical Journal of the LAOS*. 30. 271-281. 10.3233/SJI-140834.
- Kolczyńska, M. and Schoene, M. (Forthcoming) Survey Data Harmonization and the Quality of Data Documentation in T.P. Johnson, B-E Pennell, I. A. L. Stoop, & B. Dorer (eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC)*. Hoboken, New Jersey: Wiley.
- Oleksiyyenko, O., Wysmulek, I., and Vangeli, A. (Forthcoming) Identification of Processing Errors in Cross-national Surveys in T.P. Johnson, B-E Pennell, I. A. L. Stoop, & B. Dorer (eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC)*. Hoboken, New Jersey: Wiley.
- Schäfer, C., Schräpler, J. P., Müller, K.R. and Wagner, G. G. (2004) “Automatic Identification of Faked and Fraudulent Interviews in Surveys by Two Different Methods.” JSM Proceedings, Survey Research Methods Section pp. 4318–4325.
- Schoene, M. and Kolczyńska, M. (2014) “Survey Data Harmonization and the Quality of Data Documentation in Cross-National Surveys.” *CONSIRT Working Papers Series 3* at [consirt.osu.edu/](http://consirt.osu.edu/)
- Slomczynski, K.M., Powalko, P. and Krauze, T.. 2015. “Non-unique Records in International Survey Projects: The Need for Extending Data Quality Control.” *Survey Research Methods* 11(1): 1-16.
- Slomczynski, K.M., Tomescu-Dubrow, I. and Jenkins, J. C, with Kolczyńska, M., Powalko, P., Wysmulek, I., Oleksiyyenko, O., Zieliński, M.W. and Dubrow, J. (2016). *Democratic values and protest behavior in cross-national perspective. Harmonization of data from international survey projects*. Warsaw: IFiS Publishers.
- Zieliński, M., Powalko, P., and Kolczyńska, M. (Forthcoming). The Past, Present and Future of Statistical Weights in International Survey Projects: Implications for Survey Data Harmonization in T.P. Johnson, B-E Pennell, I. A. L. Stoop, & B. Dorer (eds.), *Advances in Comparative Survey Methodology: Multinational, Multiregional and Multicultural Contexts (3MC)*. Hoboken, New Jersey: Wiley.

# Surveillance, Privacy, and Leaks: Google Big Data in Cross National Perspective

by Anna H. Turner, Graduate School for Social Research, Polish Academy of Sciences

Mass surveillance, digital privacy, and data leaks have become major issues of modern society and thus major topics that social scientists study. In a technical report released in June 2017, *Public Interest in Data Surveillance and Data Leaks Before and After Snowden: Google Big Data in Cross National Perspective*, I examine the impact of Edward Snowden's revelations on public interest in surveillance, privacy, and data protection and present a methodology of using Google data in cross-national longitudinal studies. Building on the empirical literature on these topics, this project is the first to measure public interest for so many countries ( $n = 116$ ) using Google data. I use over 10,000 Google search queries and cover a period of 24 months from April 2013 (two months before the first of Snowden's revelations were published) to March 2015.

## The Snowden Data Leak

In 2013, Edward Snowden, the computer analyst and whistle blower, leaked to the press top-secret US National Security Agency (NSA) documents. Revealed in this leak was NSA's global data surveillance program, called PRISM: Since 2007, and without the knowledge or consent of their users, companies such as Google, Facebook, Microsoft and Apple had provided the NSA with their data, such as IP addresses, logins, passwords, sent messages, photos, posts, and videos. Since Snowden's disclosure, we know that the cost of using ubiquitous internet technologies is to share private data without our consent. Corporations and governments of many countries - citing safety considerations and promising various security benefits - are eager to use these data. These revelations of mass surveillance, privacy breaches, and data leaks confirm that we have a rather illusive sense of control over our data, and that those to whom we give our information have little control over its dissemination.

## Using Google Big Data

In light of these and other revelations, a major sociological topic is to monitor interest in the topics of surveillance, including the role of macro determinants that influence it. While social scientists increasingly turn to internet data, a clear methodology of how to work with these data sources is lacking. Google is the world's largest search engine and every second it processes about 55,000 queries, which rises to 4.7 billion on a daily basis. That we have access to these queries is undoubtedly of tremendous potential to social science research and an attempt to adapt the data to sociological methods is necessary.

The technical report forms the first part of my larger research project and describes in detail: (1) technical instructions about collecting the data from Google Keyword Planner; (2) methodological guidelines about preparing Google data for comparative analysis; (3) first results of the study in global regional and national perspective, where public interest in each country and continent was estimated based on number of searches for the topics of surveillance, privacy, and data protection.

*Revelations of mass surveillance, privacy breaches, and data leaks confirm that we have a rather illusive sense of control over our data.*

### Next Steps

The second part of the project is going to examine the hypothesis that in countries with a higher level of economic development (measured by GDP per capita) and democracy (measured by Democracy Index) there is greater public interest in surveillance than in other countries. A part of the impact of political and economic variables possibly goes through general and established interest in politics. This hypothesis stems from the assumption that Snowden's revelations were a sobering wake-up call, directly accusing western democratic authorities of denying their main values and undermining one of the most fundamental human rights - the right to privacy. Hence, citizens of these countries should be most concerned, as the disclosures principally affect them. For statistical analysis, we will use a longitudinal multilevel model with fixed occasions. The survey data come from the Data Harmonization project ([dataharmonization.org](http://dataharmonization.org)).

There are several methodological challenges to be addressed: First, what exactly do Google data measure? Second, creating procedures to harmonize (a) Google data - big data collected by algorithm in aggregated monthly intervals; (b) GDP per capita – public statistics collected at annual intervals; and (c) Interest in Politics - survey data collected in ad hoc time periods, grouped by a year of study.

*Anna H. Turner is a Ph.D. candidate in the Graduate School for Social Research at the Institute of Philosophy and Sociology of the Polish Academy of Sciences. She is currently completing her dissertation on understanding the role of economic, political and social macro determinants in differentiating public interest in the topics of surveillance. Her research interests include sociology of internet, information society, privacy and data protection in the context of online surveillance and methodology of research with use of Google data.*

*This project was supported by the Cross-National Studies: Interdisciplinary Research and Training Programme (CONSIRT) established by the Polish Academy of Sciences and The Ohio State University.*

*The technical report described in this article is free to download on the author's website:*

<http://annahturner.academia.edu/research>

# Conferences & Workshops

---



# seriss

SYNERGIES FOR EUROPE'S  
RESEARCH INFRASTRUCTURES  
IN THE SOCIAL SCIENCES

## **SERISS workshop “Harmonised Coding of Socio-economic Survey Questions”**

The workshop, 4th-5th September 2017 in Amsterdam, presents the work which is done in SERISS Work package 8 on “A coding module for socio-economic survey questions.” There will be presentations about the challenges of measuring and coding different socio-economic variables such as occupation, industry, employment status, educational attainment, fields of education, and social networks in surveys. The focus of the workshop is introducing a new set of tools which can be implemented in surveys for measuring the above mentioned socio-economic variables. In the workshop those tools will be presented and they can be tested in practical sessions. The workshop starts at 13:30 (registration and lunch are provided from 12:30) on 4th September and finishes at 15:00 on 5th September. For more details, visit [seriss.eu](http://seriss.eu).

## **CSDI Workshop Presentation “Applying ex-post Harmonization of Corruption Items from Cross-national Survey Projects”**

by Ilona Wyszumłek, Polish Academy of Sciences

In corruption research, surveys are among the major sources of our knowledge about the subject. However, there are several methodological challenges to studying cross-national trends in corruption with public opinion data. Corruption, given its secretive nature, is a phenomenon that is hard to capture in the interview situation. Some respondents are reluctant to answer sensitive questions about and some may understand the concept differently than intended by researchers. Moreover, international survey projects dealing with corruption continue to face challenges of unequal country representation. Estimation of rare event determinants often become problematic.

In a presentation at the Comparative Survey Design and Implementation (CSDI) workshop on March 2017 in Mannheim, Germany, I examined the methodological challenges of applying ex-post harmonization to study corruption. First, I presented the existing micro level survey data on corruption, which may allow researchers for cross-national comparisons and ex-post harmonization. I focused on corruption data from European countries during the 1989-2013 timespan, which include such projects as Eurobarometer, Life in Transition Survey, Global Corruption Barometer, Crime Victims Survey, European Social Survey, International Social Survey Program, World Values

Survey, and other - total of 63 survey waves of 19 international survey projects. I analyzed the types of gathered corruption-related questions and their frequencies of occurrence, indicating some changing trends in questionnaires. The findings showed that there is a growing interest in the subject of corruption in cross-national surveys, but the coverage of countries is uneven.

Second, based on the gathered source data, I demonstrated inequalities of representation of some European countries in the public opinion surveys with questions on corruption. I also highlighted the countries with small N problems, for which petty corruption is a rare event and the estimations of its determinants by traditional methods are imprecise. I provided an example of ex-post harmonization of bribe-giving experience variable to show which old methodological problems are solved and which new appear.

The paper is based on my dissertation project, where I study corruption experience in public schools in Europe. I follow the Survey Data Recycling framework developed by the research team lead by Kazimierz M. Slomczynski that provides a blueprint for ex-post survey data harmonization and for integrating surveys and other data sources to be used in substantive analysis. This research is supported by the Mobility Grant of the Ministry of Science and Higher Education of Poland at The Ohio State University (1292/MOB/IV/2015/0).

## Conference at Ohio State University's Mershon Center, "International Perspectives on Methods for the Study of Protest"



*Event poster picture copyright: By Jsnake17 - Own work, CC BY-SA 4.0, <https://commons.wikimedia.org/w/index.php?curid=37939878>*

Across the world, mass political protest has shaped the course of modern democracies. Building on decades of theory and methods, the two-day event, "Democracy, the State and Protest: International Perspectives on Methods for the Study of Protest," examined methodological approaches to studying the past, present, and future of democracy and protest.

It began with a conference on May 11, 2017 featuring scholars in attendance such as **Russell Dalton**, University of California, Irvine, **Jennifer Earl**, University of Arizona, **Hanspeter Kriesi**, European University Institute, Italy, **Jan van Deth**, University of Mannheim, Germany, **Patrick T.**

**Brandt**, University of Texas at Dallas, **Katerina Vrablikova**, Mershon Center OSU, **Bert Klandermans**, Free University, Amsterdam, and **Irina Tomescu-Dubrow** and **Kazimierz M. Slomczynski** from CONSIRT and the Polish Academy of Sciences.

The second day, May 12, featured a roundtable discussion with the conference speakers on the methodology of studying democracy and protest.

The event took place at the Mershon Center for International Security Studies, The Ohio State University. This event was free and open to everyone at OSU, including graduate students and other scholars from across the University.

The event was organized by the Mershon Center for International Security Studies, in collaboration with Cross-national Studies: Interdisciplinary Research and Training program (CONSIRT.osu.edu), by Irina Tomescu-Dubrow (Polish Academy of Sciences), Kazimierz M. Slomczynski (OSU and Polish Academy of Sciences), and J. Craig Jenkins (OSU).

### **Authors and Papers Delivered during the Event:**

**Russell Dalton.** *Culture, Protest and Government Performance in Established Democracies: A Reassessment of the Crisis of Democracy Thesis*

**Heidi Reynolds-Stenson and Jennifer Earl.** *The Puzzle of Protest Policing over Time: Historicizing Explanations of Social Movement Repression*

**Hanspeter Kriesi.** *European Protest in Times of Crisis: A Thirty Countries Study*

**Jan van Deth.** *Similar yet Different: Constructing Equivalent Protest Measures in Comparative Research*

**Katerina Vrablikova.** *How to Study Context and Contentious Politics? Bridging Qualitative and Quantitative Perspectives*

**Bert Klandermans and Teodora Gaidyte.** *Typologies of Political Participation: How and Why Does Political Behavior Differ across the World?*

**Patrick T. Brandt.** *Machine Learning Extensions for the Automated Coding and Categorization of Protest Events*

**Kazimierz M. Slomczynski, Irina Tomescu-Dubrow, Marta Kolczynska, Joshua K. Dubrow, and J. Craig Jenkins.** *Harmonizing Cross-national Survey Data on Participation in Demonstrations*

# Harmonization would like to hear from you!

We created this *Newsletter* to share news and help build a growing community of those who are interested in harmonizing social survey data. We invite you to contribute to this Newsletter. Here's how:

## 1. Send us content!

- Send us your announcements (100 words max.), conference and workshop summaries (500 words max.), and new publications (250 words max.) that center on survey data harmonization in the social sciences;
- Send us your short research notes and articles (500 – 1000 words) on survey data harmonization in the social sciences. We are especially interested in advancing the methodology of survey data harmonization. If we have any questions or comments about your items, we will work with you to shape them for this *Newsletter*.

Send it to: Joshua Kjerulf Dubrow, [dubrow.2@osu.edu](mailto:dubrow.2@osu.edu).

## 2. Tell your colleagues!

To help build a community, this *Newsletter* is open access. We encourage you to share it in an email, blog or social media (Facebook, Twitter, Google+, and so on).

## Support

This newsletter is a production of Cross-national Studies: Interdisciplinary Research and Training Program, of The Ohio State University (OSU) and the Polish Academy of Sciences (PAN). The catalyst for the newsletter has been the project, “Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling” (hereafter, Harmonization Project). Financed by the Polish National Science Centre in the framework of the Harmonia grant competition (2012/06/M/HS6/00322), the Harmonization Project joins the Institute of Philosophy and Sociology PAN and the OSU Mershon Center for International Security Studies in creating comparable measurements of political protest, social values, and demographics using information from well-known international survey projects. The team includes: Kazimierz M. Slomczynski (PI), J. Craig Jenkins (PI), Irina Tomescu-Dubrow (PI), Joshua Kjerulf Dubrow, Przemek Powalko, Marcin W. Zieliński, and research assistants: Marta Kolczyńska, Matthew Schoene, Ilona Wyszumlek, Olena Oleksiyenko, Anastas Vangeli, and Anna Franczak. For more information, please visit [dataharmonization.org](http://dataharmonization.org).

## Copyright Information

*Harmonization*: Newsletter on Survey Data Harmonization in the Social Sciences is copyrighted under Creative Commons Attribution-NonCommercial-ShareAlike 3.0 United States ([CC BY-NC-SA 3.0 US](https://creativecommons.org/licenses/by-nc-sa/3.0/us/)).

“You are free to: Share — copy and redistribute the material in any medium or format; Adapt — remix, transform, and build upon the material. The licensor cannot revoke these freedoms as long as you follow the license terms. Under the following terms: Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. NonCommercial — You may not use the material for commercial purposes. ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original. No additional restrictions — You may not apply legal terms or technological measures that legally restrict others from doing anything the license permits.”