Cross-national Studies: Interdisciplinary Research and Training Program (CONSIRT), The Ohio State University and the Polish Academy of Sciences

Harmonization:
Newsletter on Survey Data
Harmonization in the Social Sciences

Volume 2, Issue 1
Summer 2016
Editors
Irina Tomescu-Dubrow
and
Joshua Kjerulf Dubrow
CONSIRT

consirt.osu.edu/newsletter/
ISSN 2392-0858

Working Together

Welcome to the third issue of *Harmonization*: Newsletter on Survey Data Harmonization in the Social Sciences. Survey data harmonization and big data are innovative forces in the social sciences. Working together, we share news and communicate with the growing community of scholars, institutions and government agencies who work on harmonizing social survey data and other projects with similar focus.

This issue features articles on the issues of data quality, duplicate cases, and survey weights. In "Quality of Survey Data," Revilla, Saris, and colleagues, present the Survey Quality Predictor (SQP) as a way to account for measurement error in surveys. SQP can be used to design survey items or correct for measurement errors after the survey has been collected.

Sarracino and Mikucka's article, "Estimation Bias Due to Duplicated Observations," uses a Monte Carlo simulation to understand how duplicate records impact estimates and evaluates the effectiveness of some solutions. They evaluate whether "the risk of obtaining biased estimates of regression coefficients increases with the number of duplicate records." This article summarizes some of their ongoing research on duplicates.

Finally, Kołczyńska and colleagues' article, "Survey Weights," is in the context of the growing popularity of weighting data as means to contend with sampling and non-responses errors. They propose that properties of weights could be used to evaluate the quality of weights, and as indicators of the quality of the data as a whole.

In this newsletter, we also present news about GESIS's CharmStats, a research grant from The Ohio State University's Mershon Center, a report on the recent data harmonization conference held in Warsaw in December of last year, and an abstract of a presentation at the International Sociological Association Forum of Sociology in Vienna, Austria, 2016.

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

Acknowledgements

Editors thank Marta Kołczyńska for technical assistance.

Articles

Quality of Survey Data: How to Estimate It and Why It Matters

by Melanie Revilla, Willem Saris and the Survey Quality Predictor (SQP) team

There is no measurement without error. However, the size of the error can vary depending on the measure used. In particular in social sciences survey data, the size of the error can be very large: on average, 50 percent of the observed variance in answers to survey questions is error (Alwin 2007). The size of the error can vary a lot depending on the exact formulation of the survey questions used to measure the concepts of interest (Saris and Gallhofer 2014) and also across languages or across time. Thus, one of the main challenges for cross-sectional and longitudinal surveys, in order to make meaningful comparisons across groups or time, is to be able to estimate the size of the measurement error, and to correct for it.

SQP is based on 3,700 quality estimates of questions obtained in more than 30 European countries and languages…

One way to estimate the size of the measurement errors (both random and systematic errors) is the multitrait-multimethod (MTMM) approach, first proposed by Campbell and Fiske (1959), and then developed further by many authors (in particular, Jöreskog 1970; Andrews 1984; Saris and Andrews 1991). In this approach, different questions (called "traits") need to be repeated using different methods (for example, a 5-point scale, a 7-point scale and an 11-point scale). Usually, for identification purposes, at least three different traits are measured using at least three different methods. Under quite general conditions (cf. Saris and Andrews 1991), this approach allows estimating the reliability (1- random errors variance) and validity (1 - method error variance) of a set of survey questions. By taking the product of these reliability and validity, we get an estimate of the total quality of a survey question, which can also be defined as the strength of the relationship between the concept one is really interested in and the observed answers. The closer to 1 this quality estimate is, the lower the level of measurement errors for a given question.

One of the main limits of the MTMM approach lies in the necessity of repeating similar questions to the same respondents, which can lead to cognitive burden, higher cost, longer surveys, etc. Moreover, the results from the MTMM analyses are specific to the questions included. It is not possible to generalize from these questions to all the other questions in the survey; though, it is also not possible to repeat all the survey questions, which would be like asking respondents to complete twice the same survey in a row.

Therefore, Saris and Gallhofer (2014) proposed an alternative: use cumulative data from past MTMM experiments for a meta-analysis and investigate the effect of questions' characteristics and questions' context on the reliability and validity. Then, use this information to predict the quality of new survey questions based on their own characteristics. This is what the Survey Quality Predictor (SQP) software does in a user-friendly way.

SQP is a survey quality prediction system for questions used in survey research and a database of questions with information about their quality. The software is available for free at sqp.upf.edu. SQP is based on 3,700 quality estimates of questions obtained in more than 30 European countries and languages by MTMM analyses using the True Score model proposed by Saris and Andrews (1991). Most of these MTMM experiments have been done in the European Social Survey (ESS). Indeed, in each ESS round, four to six MTMM experiments are included in almost all the participating countries. In each experiment, three traits are measured using three or four different methods. SQP provides the users the possibility to consult the MTMM estimates for all these questions and languages. In addition, the program makes predictions of the quality of new questions on the basis of information about the choices that have been made with respect to the questions' characteristics. The user needs to code the characteristics of his/her questions, and in that way, can get a prediction of the quality, without needing to collect any new data. Some brief tutorials explaining what SQP is and how it works are available at: https://www.youtube.com/channel/UCpljiQFlE4j5CYI-rqMKDig

The information from SQP or from the MTMM experiments can be used in different ways. In particular, it can be used before data collection to help designing the questions (Revilla, Zavala and Saris 2016), and after data collection in order to correct for measurement errors (De Castellarnau and Saris 2014; Saris and Revilla 2016). These are two crucial steps in order to get proper estimates of the substantive relationships of interest. However, even if the tools are available, in practice, these techniques are not implemented by most researchers. We believe that for the future of survey research, this issue needs to be given more attention.

References

Alwin, D.F. (2007). *Margins of error: A study of reliability in survey measurement*. Hoboken, Wiley.

Andrews, F. (1984). Construct validity and error components of survey measures: A structural modeling approach. *Public Opinion Quarterly*, 46:409–442.

Campbell, D. T., and D. W. Fiske (1959). Convergent and discriminant validation by the multitrait-multimethod matrices. *Psychological Bulletin*, 56, 81-105.

De Castellarnau, A. and Saris, W. E. (2014). A simple way to correct for measurement errors. European Social Survey Education Net (ESS EduNet). Available at: http://essedunet.nsd.uib.no/cms/topics/measurement/

Jöreskog, K.G. (1970). A general method for the analysis of covariance structures. *Biometrika*, 57:239–251.

Revilla, M., Zavala Rojas, D., and W.E. Saris (2016). "Creating a good question: How to use cumulative experience". In Christof Wolf, Dominique Joye, Tom W. Smith and Yang-Chih Fu (editors), *The SAGE-Handbook of Survey Methodology*. SAGE.

Saris, W.E. and F.M. Andrews (1991). Evaluation of measurement instruments using a structural modeling approach. In P. Biemer, R.M. Groves, L. Lyberg, N. Mathiowetz, S. Sudman (Eds.), *Measurement errors in surveys* (pp. 575-597). New York: Wiley.

Saris, W.E. and I.N. Gallhofer (2014). *Design, evaluation and analysis of questionnaires for Survey research*, Hoboken, Wiley (second edition).

Saris, W.E., and M. Revilla (2016). "Correction for measurement errors in survey research: necessary and possible". *Social Indicators Research,* 127(3): 1005-1020. First published online: 17 June 2015. DOI: 10.1007/s11205-015-1002-x

Melanie Revilla is researcher at the Research and Expertise Centre for Survey Methodology (RECSM) and adjunct professor at Universitat Pompeu Fabra (UPF, Barcelona, Spain).

Willem E. Saris is Professor and researcher at the Research and Expertise Centre for Survey Methodology of the Universitat Pompeu Fabra (Spain). Together with Daniel Oberski, he was awarded the AAPOR Warren J. Mitofsky Innovators Award 2014 for the Survey Quality Predictor (SQP 2.0).

Estimation Bias due to Duplicated Observations: A Monte Carlo Simulation

by Francesco Sarracino and Małgorzata Mikucka

Two recent, independent studies documented that duplicate records are frequent in many international surveys (Kuriakose and Robbins 2015; Slomczynski et al. 2015). Yet, the literature neglects the influence that duplicate records may have on the analysis of statistical data. As surveys are an important source of information for modern societies, filling this gap is a sensible task. Using a Monte Carlo simulation, we found that duplicate records create considerable risk of obtaining biased estimates (Sarracino and Mikucka 2016). For instance, if the dataset contains about 10 percent of duplicated observations, then the probability of obtaining correct estimates is about 11 percent. Weighting the duplicate cases by the inverse of their multiplicity is the method that minimizes the possibility of errors when multiple doublets are present. These findings call for further research on strategies to analyze affected data, and ask for more care in data collection procedures.

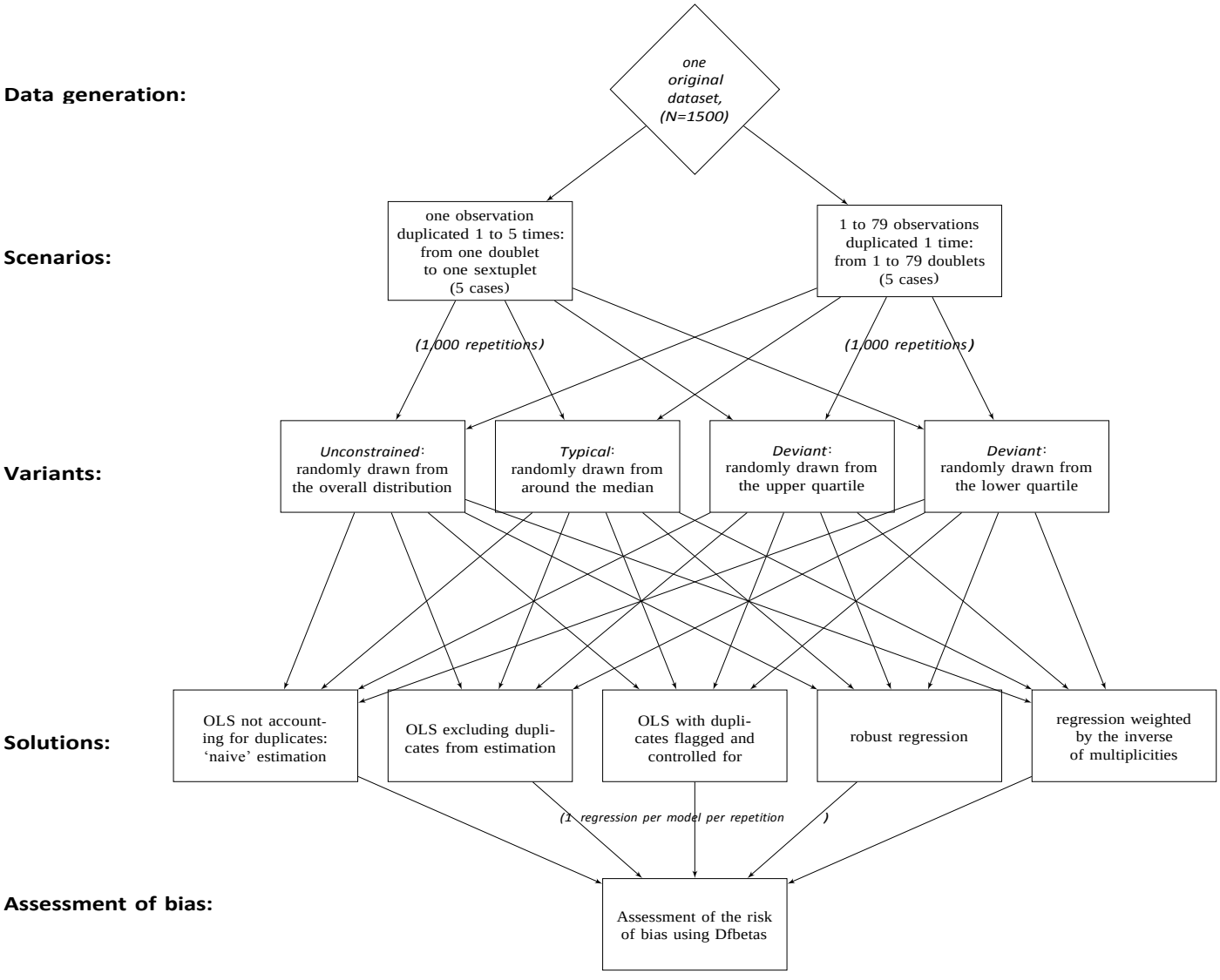
The work of applied researchers often relies on survey data, and the reliability of the results depends on accurate recording of respondents' answers. Yet, sometimes this condition is not met. A recent study by Slomczynski et al. (2015) investigated survey projects that are widely used in the social sciences and reported a considerable number of duplicate records in 17 out of 22 projects. Duplicate records in surveys are records that are not unique, that is, records in which the set of all (or nearly all) answers from a given respondent is identical to that of another respondent.

The causes and the methods to identify duplicate records are a source of fierce debate. Yet, it seems that scholars agree that whatever the conclusion, duplicate records will remain and social scientists need to find a way to deal with them. This is the aim of our recent work (Sarracino and

Mikucka 2016): studying how duplicate records affect estimates in the social sciences and evaluating the effectiveness of some of the possible solutions. In particular, we consider the following solutions: excluding the duplicate cases from the analysis, flagging the duplicate cases and including the flags in the model, using robust regression model as a way to minimize the effect of influential observations, and weighting the duplicate cases by the inverse of their multiplicity.

To this aim, we use a Monte Carlo simulation. Our analysis consists of four main steps. In the first step we generate the initial dataset. In the second step we duplicate randomly selected cases according to the two scenarios and four variants mentioned above. In the third step we estimate regression models using a 'naive' approach, i.e. treating data with duplicates as if they were correct; we also estimate regression models using five possible solutions to deal with duplicate cases. Finally, we compare the bias of estimates obtained from various scenarios of cases' duplication and we evaluate the effectiveness of the possible solutions. Figure 1 provides an overview of our strategy.

Figure 1: Diagram Summarizing the Empirical Strategy



Results show that the risk of obtaining biased estimates of regression coefficients increases with the number of duplicate records. If data include less than 1 percent of duplicate records, the probability of obtaining unbiased estimates is 41.6 percent. If duplicate records sum up to about 10 percent of the sample, the probability of obtaining unbiased estimates reduces to about 11.4 percent. These figures do not change significantly if the duplicate records are concentrated around the mean or on the ties of the distribution of a variable.

To date, practitioners have limited tools to deal with the issue of duplicate records.

In sum, our results indicate that even a small number of duplicate records create considerable risk of obtaining biased estimates. This suggests that practitioners who fail to account for the presence of duplicate records face a considerable risk to reach misleading conclusions.

Unfortunately, to date, practitioners have limited tools to deal with the issue of duplicate records. The first best is, of course, not having duplicates at all – which calls for putting more care in the data producing phases. Yet, when the duplicate records are in the data, little can be done to minimize their effects. Among the five solutions we considered, weighting the duplicates by the inverse of their multiplicity provided the most encouraging results. This solution outperforms 'naive' estimates in presence of one doublet, and it performs equally to dropping or flagging the duplicates when one triplet, quadruplet, quintuplet or sixtuplet are present in the data. However, the performance of this solution decreases when the number of duplicates increases, yet the chances of error-free estimates are higher than in the alternative solutions.

Our results are discouraging, but not pessimistic: although duplicate data plague some of the major surveys currently used in social sciences, it is possible to limit the risk of biased estimates. Yet, the best solution is preventing the presence of duplicate records since data correction with statistical tools is not a trivial task. This calls for further research about how to address the presence of multiple doublets in the data and more refined statistical tools to minimize the consequent estimation bias.

References

Kuriakose, N. and Robbins, M. (2015). Falsification in surveys: Detecting near duplicate observations. *Available at SSRN*. Accessed on 28th of July 2015.

Sarracino, F. and Mikucka, M. (2016). Estimation bias due to duplicated observations: a monte carlo simulation. MPRA Paper No. 69064, University Library of Munich, Germany.

Slomczynski, K. M., Powalko, P., and Krauze, T. (2015). The large number of duplicate records in international survey projects: The need for data quality control. *CONSIRT Working Papers Series 8 at*

consirt.osu.edu.

Francesco Sarracino is a researcher at the Institut national de la statistique et des études économiques du Grand-Duché du Luxembourg (STATEC), and LCSR National Research University Higher School of Economics, Russian Federation. 13, rue Erasme, L-2013, Luxembourg. Tel.: +352 247 88469; e-mail: f.sarracino@gmail.com

Małgorzata Mikucka is a researcher at the Centre for Demographic Research, Université Catholique de Louvain, Belgium.

Survey Weights as Indicators of Data Quality

by Marta Kołczyńska, Marcin W. Zieliński, and Przemek Powałko

In the last decades, more and more scholars are using weights as a procedure for correction of distortions in surveys. The improvement in the quality of the data using weights is conditional upon the quality of the weights themselves, as well as their ability to correct the discrepancies between the realized sample and the population. In cross-national research, especially when combining survey data from different survey projects, the additional challenge is making sure across national samples, the quality of the weights and the quality of weighted data are comparable and allow for meaningful analyses of the combined data.

Over time, weighting data has gained popularity as a way of dealing with sampling and non-responses errors.

We propose four properties of weights that can be considered as both indicators of their quality, and also as indicators of the quality of the data in terms of the degree of distortion between the targeted sample and the achieved sample. First, the mean value of weights in a sample should be equal to 1; otherwise weighting the data would change the sample size and thus artificially alter standard errors and confidence intervals and lead to unfounded conclusions of hypothesis testing. Second, while weights usually lead to an increase in variance in the data, weights with a smaller variance are generally preferred over weights with greater variance. Weight variance depends on the discrepancy between the achieved sample and the population, or the extent to which the raw data need to be corrected to represent the population. Thus, in some sense, the weight variance can be assumed as a rough indicator of the quality of the sample. Finally, to avoid case exclusion

and the loss of information, weights should have values greater than 0. If a weight would take the value 0, that case would be excluded from analyses. Extreme values should be avoided because they lead to potential bias if the individuals who have been assigned very high weights are specific, unusual, and deviating from the average.

We present some of our findings from the analysis of these properties, as well as mean variance, minimum and maximum value, of weights in 22 international survey projects conducted between 1966 and 2013 in 142 countries and territories. The database contains 1,721 national surveys; for detailed information about the project see dataharmonization.org. To start, we analyzed the availability of weights in our database and found them in 60.2 percent (n=1,035) of all surveys. Over time, weighting data has gained popularity as a way of dealing with sampling and non-responses errors. The main increase in the adoption of weights happened in late 1990s and early 2000s. In general, it is clear that the availability of weights has been increasing with time.

Mean

Of all samples in the Harmonization project that have weights, 70 percent do not meet the criterion of the mean of weights equal to 1. Since small deviations of the mean from 1 could be the result of rounding, we also took a less strict approach and considered weights to be correct if the deviation was less than 0.001. Using this definition, we find that of the 22 projects, only five have all surveys with correctly calculated means: AMB, CNEP, EQLS, ESS and WVS. On the other hand, only 20 percent of surveys in VPCPCE and 27 percent in NBB have weights that meet this criterion. When looking at surveys from all projects together, in 12.7 percent of national surveys the mean of weights exceeds the specified range of 0.999 to 1.001, with values ranging from 0.83 (Philippines, ASB 2010) to 3.29 (Philippines, ISSP 1996).

Standard Deviation

There is no set value for the highest acceptable value of standard deviation (i.e. benchmarks do not exist). In practice, and from the perspective of data weighting, we argue that the lower the standard deviation, the better. Our analysis shows that standard deviation ranges from 0.00 (Panama, Latinobarometro, 2002) to 2.85 (New Zealand, ISSP, 2007 with an average value of 0.43.

Interestingly, we found an association between the standard deviation of weights and the population density. This means that in countries where the population density is lower, the level of discrepancy between sampled (and thus between population) and achieved samples is bigger. This can indicate some problems with reaching some specific groups of respondents in these countries and that this process is not random in terms of factors taken into account in sampling procedure.

Extreme Values

We found that minimum values of weights range from exactly 0 (in 42 surveys) to 1.91 (Philippines, ISSP, 1991). The lowest maximum value of weights, 0.92, was found in the New

Baltic Barometer survey in Lithuania, 2001. The highest, 90.32, in New Zealand, ISSP 2007. The issue of cases multiplied by weighting is very similar in its effects to the problem of outliers (Benmei 2004), even though the source in the former case is artificial. One of the potential solutions to the problem of high values of the maximum and minimum in calculated weights is trimming extreme weights (Biemer 2010; Little et al. 1997).

Conclusion

Weights can also be used to evaluate the quality of survey data. For example, excluding cases by assigning zero weights suggests errors in weight computation or deficiencies in the documentation provided by the authors of the studies. Errors related to weights should be analysed together with other types of errors, especially if they are to be used as a source of meta-data information about the overall survey quality. It is important to keep in mind that correcting formal errors in weights, e.g. rescaling weights whose mean differs from 1 back to the original sample size, improves the quality of weights in the technical sense, but does not ensure that the weights have been calculated properly, or that using them as correction factors improves the representativeness of the sample.

Table 1. Quality of Survey Weights by Survey Project

Project (1)	% correct weights* (2)	Mean			Standard deviation			Minimum		Maximum	
		Mean (3)	Lowest (4)	Highest (5)	Mean (6)	Lowest (7)	Highest (8)	Lowest (9)	Highest (10)	Lowest (11)	Highest (12)
AFB	88.1	1.00	0.89	1.03	0.39	0.08	1.03	0.01	1.00**	1.01	7.51
AMB	100.0	1.00	1.00	1.00	0.47	0.05	0.68	0.00	0.91	1.05	4.81
ARB	80.0	1.00	1.00	1.01	0.45	0.15	1.05	0.05	0.72	1.54	16.48
ASB	76.0	0.99	0.83	1.00	0.40	0.13	0.92	0.03	0.80	1.19	12.76
ASES	***	***	***	***	***	***	***	***	***	***	***
CB	***	***	***	***	***	***	***	***	***	***	***
CDCEE	75.0	1.00	1.00	1.02	0.44	0.11	1.10	0.03	0.80	1.3	8.52
CNEP	100.0	1.00	1.00	1.00**	0.40	0.21	0.63	0.21	0.53	1.67	3.16
EB	96.2	1.00	0.97	1.02	0.34	0.03	1.14	0.00	0.91	1.07	6.23
EQLS	100.0	1.00	1.00	1.00	0.53	0.27	1.21	0.02	0.62	1.67	12.86
ESS	100.0	1.00	1.00	1.00	0.46	0.01	1.08	0.00	0.98	1.02	5.98
EVS	74.2	1.00	0.97	1.01	0.36	0.09	0.97	0.00**	0.88	1.17	18.18
ISJP	88.9	1.01	1.00	1.09	0.36	0.01	0.64	0.03	1.00**	1.27	6.52
ISSP	75.9	1.02	0.88	3.29	0.52	0.10	2.85	0.00**	1.91	1.13	90.32
LB	88.7	1.00	0.98	1.05	0.44	0.00	1.50	0.00**	1.00**	1.00**	19.66
LITS	96.9	1.00	1.00	1.00	0.37	0.14	0.63	0.00	0.75	1.11	4.53
NBB	27.8	1.02	0.89	1.49	0.27	0.03	0.92	0.08	0.97	0.92	4.87
PA2	***	***	***	***	***	***	***	***	***	***	***
PA8NS	***	***	***	***	***	***	***	***	***	***	***
PPE7N	33.3	1.07	1.00	1.21	0.53	0.28	0.89	0.28	1.00**	1.2	3.97
VPCPCE	20.0	1.00	0.98	1.02	0.30	0.10	0.49	0.41	0.68	1.18	3.23
WVS	100.0	1.00**	1.00**	1.00**	0.42	0.02	2.22	0.04	0.97	1.03	32.25

This work is funded by the project conducted by the Institute of Philosophy and Sociology of the Polish Academy of Sciences and the Mershon Center for International Security Studies at The Ohio State University, USA with funding from the National Science Centre, Poland (grant number 2012/06/M/HS6/00322).

Marta Kołczyńska is a Ph.D. candidate at the Department of Sociology, The Ohio State University, and a research assistant in the Harmonization Project.

Marcin W. Zieliński, Ph.D., is Assistant Professor at the Institute of Philosophy and Sociology, Polish Academy of Sciences and specialist at the Robert B. Zajonc Institute for Social Studies, University of Warsaw which is affiliated with the Institute for Social Research at the University of Michigan. He is co-author and member of the team conducting Polish General Social Survey (PGSS) and the Principal Investigator of the Polish edition of the International Social Survey Programme (ISSP). He is also the head of the Polish Social Data Archive (ADS).

Przemek Powałko currently works as a database specialist for the Data Harmonization Project at the Institute of Philosophy and Sociology, Polish Academy of Sciences.

News

Publication: New Standards Proposed for Documenting Variable Harmonization

by Kristi Winters

An online article proposing standards for the documentation, publication and scholarly citation of variable harmonization appeared in the open access journal *PLoS One* earlier this year: "Proposed Standards for Variable Harmonization Documentation and Referencing: A Case Study Using QuickCharmStats 1.1", written by Dr. Kristi Winters and Dr. Sebastian Netscher (both from GESIS), identifies the necessary and sufficient information required to replicate a variable harmonization.

The study level, question level and variable level metadata are recommended as required information. The documentation can be built digitally using the free and open-source harmonization software, CharmStats, or filled in by hand. After collating this information researchers can deposit their documentation for free in the online repository datorium, provided and maintained by GESIS, after the submission has undergone peer review. Once accepted, each submission is provided with its own DOI and hyperlink. Researchers can then cite their

harmonization documentation in articles or other publications rather than use the word count to describe what they did with their data.

For more information on how to submit a variable harmonization for peer review and publication with datorium or to download and use CharmStats products, email charmstats@gesis.org.

The Ohio State University's Mershon Center Funds New Harmonization Endeavour on Democracy and Protest

Researchers from The Ohio State University and the Polish Academy of Sciences have been awarded a grant for new harmonization work that will aid building and testing theories of democracy and protest.

The project "State Responses to Contention and New Waves of Protests: Using Harmonized Survey Data for 82 Countries" (academic year 2016-2017) is financed by The Ohio State University's Mershon Center for International Security Studies, through a Faculty Research and Seed Grant. The Principal Investigators are Kazimierz M. Slomczynski, J. Craig Jenkins, and Irina Tomescu-Dubrow.

This project will add substantive knowledge on the relationship between democracy and political protest in comparative perspective, while also advancing the methodology of cross-national survey data analysis.

A main contribution lies in overcoming unequal country representation in international survey projects by integrating extant projects and harmonizing selected variables ex post. In doing so, the researchers will draw on, and expand, the Democratic Values and Protest Behavior project (dataharmonization.org), conducted by CONSIRT. Specifically, they will enrich the database of 1721 national surveys selected from 22 international projects, including the ESS, ISSP, World Values Survey, regional barometers, and specialized surveys on political behavior, with (a) new contextual indicators and (b) new harmonized variables. They will use these data to explain individual participation in protest – conventional, disruptive and/or violent – as a linear and interactive function of peoples' democratic values, their trust in public institutions, a set of socio-demographics, and various country-level characteristics, such as economic development, democratic practices and mobilization resources.

As part of this project, the researchers will produce (a) the project dataset, which will be publicly archived; (b) research papers to be submitted for publication consideration in top ranked social science journals, and (c) an international conference and workshop on "Democracy, the State and Protest: International Perspectives on Methods for the Study of Protest," to be held in May 2017 at the OSU Mershon Center for International Security Studies.

Report on Event: "Longitudinal Survey Research: Methodological Challenges"

by Joshua Kjerulf Dubrow and Irina Tomescu-Dubrow, CONSIRT and IFiS PAN

Cross-national Studies: Interdisciplinary Research and Training program (CONSIRT.osu.edu) organized the event, "Longitudinal Survey Research: Methodological Challenges," on December 15-18, 2015, at the Institute of Philosophy and Sociology, Polish Academy of Sciences (IFiS PAN), in Warsaw, Poland.

The common theme of the Warsaw international event was methodological challenges in cross-sectional time series and panel surveys. These types of data have been crucial to generating key insights into the conditions, causes and consequences of social change. Ironically, the very change that social scientists examine – technological, economic, political and cultural – poses serious threats to traditional survey methods. New communication modes, declining response rates worldwide, the spectacular growth of big data from non-survey sources and their increasing popularity in the social sciences, constitute such threats. Survey administrators are forced to re-think their methods, from how to design surveys, contact respondents, and ask questions, to how to analyze, store, and distribute the data.

Threats are accompanied by opportunities. The event discussed how advances in both survey methods and communication and computational technologies, combined with the rise of interdisciplinary collaborative scientific teams and laboratories across the social sciences, can aid social science methodology and provide new substantive insights.

The event was composed of two parts. First was the conference, "The Present and Future of Longitudinal Cross-sectional and Panel Survey Research" (December 15-16). Its purpose was to engage established scholars, young researchers, and graduate students from different countries and disciplines, in discussions of the present and future of longitudinal surveys. Day One of the conference featured two sessions, the first devoted to international cross-sectional surveys, and the other to panel surveys. Key questions for both sessions included:

A. What are the most troublesome methodological challenges that major longitudinal surveys face now, and in the next ten years? How can these challenges be met, and overcome?

B. To improve data quality, should we standardize survey documentation across international survey projects, beginning with guidelines provided by the Data Documentation Initiative (DDI)? If so, how can this be achieved?

C. What are speakers' visions of the future of survey methodology – from survey design to data access and storage – for the next wave, and for the next ten years?

Christof Wolf, GESIS, Germany, delivered the Plenary Lecture: "Challenges of Survey Research." It was followed by Session One, "Longitudinal Cross-sectional Survey Research," with **Christian Welzel**, Leuphana University of Lueneburg, Germany, as discussant. Among presenters we welcomed **Rory Fitzgerald**, City University London, UK, who presented "Facing Up to the Challenges and Future of Repeat Cross-sectional, Cross-national Social Surveys. The Synergies for Europe's Research Infrastructures in the Social Sciences Initiative;" **Melanie Revilla**, Pompeu Fabra University, Spain, on "Quality of Survey Data: How to Estimate It and Why It Matters;" **Peter Granda**, University of Michigan and ICPSR USA, on "Survey Data Documentation: The Disjunction between Description and Assessment;" and **Mitchell Seligson**, LAPOP, Vanderbilt University USA, on "The AmericasBarometer by LAPOP: Challenges in Cross-National Longitudinal Surveys."

The second session was on Panel Survey Research, with **Dean Lillard**, The Ohio State University USA and chief of the CNEF harmonized panel survey project, as the discussant. We welcomed two presenters: **Elizabeth Cooksey**, NLSY, The Ohio State University USA, on "Methodological Challenges in the US National Longitudinal Surveys of Youth" and **Oliver Lipps**, FORS, Switzerland, on "Methodological Challenges of Panel Surveys Now and in Ten Years – A Swiss Perspective."

Day Two of the conference "POLPAN: Preparing for the First 30 Years" focused on the Polish Panel Survey, POLPAN 1988 – 2013. POLPAN is the longest running panel survey conducted on a national representative sample of individuals in Central and Eastern Europe. Preparation for the 2018 wave just begins. In Session One, **Kazimierz M. Slomczynski** and **Zbigniew Sawiński**, who have led POLPAN over the decades, discussed how POLPAN dealt with the difficult questions Day One posed. In Session Two the presenters focused on POLPAN's future, including its tremendous relevance for research on social structure. Elizabeth Cooksey chaired the session.

The afternoon of Day Two focused on empirical findings from the POLPAN data, including the 2013 wave. We welcomed the following presentations: **Małgorzata Mikucka**, University of Leuven, Belgium, on "What Affects Subjective Evaluation of Health?"; **Zbigniew Karpiński**, IFiS PAN, and **Kinga Wysieńska-Di Carlo**, Albert Shanker Institute USA, and IFiS PAN, on "Applying Survival Analysis to Understand the Motherhood Penalty in a Dynamic Framework"; and **Anna Kiersztyn**, University of Warsaw, Poland, "Over-education in Poland, 1988-2013: Driving Factors and Consequences for Workers."

The workshop "Harmonization of Survey and Non-Survey Data" (December 17-18) constituted the second part of the December international event in Warsaw. This workshop was devoted to issues of *ex post* harmonization of survey data in the context of the Harmonization and Survey Data Recycling projects.

The first day of the workshop focused on harmonization of international survey projects. We discussed the concept of survey data recycling (SDR) as a new way of reprocessing information from extant cross-national projects in ways that minimize the "messiness" of data

built into original surveys, that expand the range of possible comparisons over time and across countries, and that improve confidence in substantive results. The workshop highlighted various steps of SDR via examples of substantive target variables that we created using information from well-known international survey projects (e.g. WVS, ISSP, ESS, various regional barometers).

Kazimierz M. Slomczynski and **Irina Tomescu-Dubrow** started the session with an overview of the Harmonization Project. It was followed by two Round-table Discussions on the topics of "Presenting, Storing and Accessing Information on Source Variables" and "Quality of Data and Harmonization Processes," respectively. We learned from, and enjoyed, the spirited discussion led by **Dean Lillard**, The Ohio State University, USA, **Christof Wolf**, GESIS, **Peter Granda**, University of Michigan and ICPSR, USA, **Mitchell Seligson**, LAPOP, Vanderbilt University, USA and **Markus Quandt**, GESIS.

Day Two of the workshop assessed possibilities of harmonizing longitudinal survey data with the East European Parliamentarian and Candidate data (EAST PaC), with a focus on women's political inequality. EAST PaC consists of all candidates who stood for national parliamentary elections in Poland, Hungary and Ukraine from the 1990s to the 2010s. Candidates are matched over time. This renders a dataset that allows researchers to track the political careers of every candidate, from the thousands who never won to the few political lifers whose parliamentary careers span decades. **Joshua K. Dubrow** presented an overview of the Electoral Control project and the uses of EAST PaC data. Participants evaluated opportunities of jointly using these data with POLPAN and other surveys. We engaged in an extended discussion on improving our knowledge, via survey data and non-survey data sources, on gender and values worldwide. **Amy C. Alexander**, Quality of Government Institute Sweden, **Catherine Bolzendahl**, University of California-Irvine USA, and **Tiffany Barnes**, University of Kentucky, USA, led this discussion.

This international event was funded by several grants from Poland's National Science Centre, including, "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling," in the framework of the Harmonia grant competition (2012/06/M/HS6/00322); Polish Panel Survey, POLPAN 1988-2013: Social Structure and Mobility (2011/02/A/HS6/00238); and "Who Wins and Who Loses in the Parliamentary Elections? From Formal Theory to Empirical Analysis," (Sonata Bis decision number 2012/05/E/HS6/03556). The event was also supported by the Institute of Philosophy and Sociology, Polish Academy of Sciences.

Presentation at the 3rd ISA Forum of Sociology in Vienna, 10-16 July 2016: "Linking National Surveys, Administrative Records and Mass Media Content: Methodological Issues of Constructing the Harmonized Data-File"

by Ilona Wysmułek, Olena Oleksiyenko, Przemek Powałko, Marta Kołczyńska, Marcin W. Zieliński, Kazimierz M. Slomczynski, and Irina Tomescu-Dubrow

In the presentation, we discuss the opportunities of construction of the harmonized data-file that links data from three sources: national surveys, administrative records, and the media. The basis of the data-file comes from 22 well-known international survey projects containing questions on protest behavior, which consists of 1721 national surveys covering 132 countries. The data from administrative country-level records on population size, ethnic fractionalization, GDP and other characteristics, as well as media content (e.g. event data on protest) are incorporated into the integrated data-file. From the methodological point of view, there are a number of challenges to overcome for reaching the aim of the project: building the integrated data-file. In the presentation we concentrate on proposed ways of linking data for multi-level analyses, with countries and years as macro-levels. We discuss data quality on both the micro- and macro-levels, and some aspects of secondary data usage of survey and non-survey data together. The logic of data linkage and data processing procedures are of general nature and can be applied to other comparative projects. The paper is a part of the project "Democratic Values and Protest Behavior: Data Harmonization, Measurement Comparability, and Multi-Level Modeling in Cross-National Perspective", financed by the Polish National Science Centre (2012/06/M/HS6/00322), located at the Polish Academy of Sciences and The Ohio University.

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.

2. Tell your colleagues!

- To help build a community, this *Newsletter* is open access.

- We encourage you to share this newsletter 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 is our ongoing 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, Joshua Kjerulf Dubrow, Przemek Powałko, Marcin W. Zieliński, and research assistants: Marta Kołczyńska, Matthew Schoene, Ilona Wysmułek, Olena Oleksiyenko, Anastas Vangeli, and Anna Franczak. For more information, please visit dataharmonization.org.

Copyright Information