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 5, Issue 2
Fall 2019

Editors
Irina Tomescu-Dubrow
and
Joshua K. Dubrow
CONSIRT

consirt.osu.edu/newsletter
ISSN 2392-0858

Eyes on the Horizon

Welcome to *Harmonization: Newsletter on Survey Data Harmonization in the Social Sciences*. This newsletter celebrates and promotes the community of scholars, institutions, and government agencies that produce fascinating harmonization research. With eyes on the horizon of rigorous scholarship, creative exploration, and fruitful intellectual debate, the community and this newsletter move forward the interdisciplinary methodological field of data harmonization.

This issue features a diverse set of articles and the latest community news. The first article, by **Dominique Joye**, **Marlène Sapin** and **Christof Wolf**, is an open call for more transparency and research on weights in cross-national surveys. Then, **Patricia Hadler** and **Agnès Parent-Thirion** argue on why it is a good idea to use cognitive interviews and "web probing" for cross-cultural cognitive pretesting. Rounding out the article section is **Bogdan Voicu**'s exploration of how events during fieldwork periods can influence survey comparability. For community news, we present a series of reports: **Andrew James Klassen** introduces his new large-scale survey data harmonization project called "HUMAN Surveys," **Dara O'Neill** and **Rebecca Hardy** report on the CLOSER Consortium for UK longitudinal surveys, and **Sarah Martin** discusses the harmonization of surveys of Scotland. Finally, the Principal Investigators of two recent Polish National Science Foundation grants on harmonization, **Piotr Jabkowski** and **Piotr Cichocki**, and **Anna Kiersztyn**, report on their projects.

As with every issue of *Harmonization*, we welcome your articles and news. Please send them to the newsletter editors at tomescu.1@osu.edu and dubrow.2@osu.edu.

In This Issue…

The editors thank Ilona Wysmulek, Joonghyun Kwak, and Kazimierz M. Slomczynski for their assistance with this newsletter.

Articles

Weights in Comparative Surveys? A Call for Opening the Black Box

by Dominique Joye, Marlène Sapin, and Christof Wolf

Weights in survey research seem to be a disputed object: some surveys' producers insist on the use of weights (Lavallée & Beaumont, 2015) even if the discussion on the construction and impact of these weights is sometimes not fully explicit (Gelman, 2007a). Along the same lines, if weights are used, there is some fear that they will modify the results "too much." The idea of trimming or reducing the range of weights could be seen as an indication of taking this direction, even if the statistical reasoning is different (Potter & Zeng, 2015). Last but not least, while the discussion seems rather vivid between statisticians (Gelman, 2007 and comments in *Statistical Science*, Vol. 22 (2); Little, 2012), it seems far less developed from a social sciences perspective.

This last point is quite interesting, specifically when a renowned statistician like Kish (1994) insists on the difference between design and analysis – the latter probably being more familiar to users of surveys. Kish (1994) writes,

> "Weighting belongs to analysis, and its most important forms have much in common with substantive analysis. It represents a form of control by removing disturbing variables from those variables that should belong to the definitions of the predictor variable. For example, should the mean family income in multinational (or domain) comparisons be adjusted for cost of living, and/or for sizes of families, and/or for urban/rural differences? However, some other aspects of weighting are related to sample design and belong to statistical analysis. For example, some weighting may increase variances, hence may perhaps be foregone for lower mean square errors. Weighting for unequal selection probabilities relates to the sample design" (p. 173).

In this paper, we take the position of social scientists interested in promoting a sensible use of international surveys and in opening a debate on a topic too often confined in informal exchanges between colleagues. This is even more important when observing a radical change in the way surveys are organised in different countries: change of modes, change in response rates, and an increase in data linking and in the combination of information sources (Joye et al., 2016). Of course, we do not have room to discuss all these elements, but we will insist that weighting is even more complicated in such a complex environment and that it is related to all steps of the survey life-cycle.

Let us examine a few points in regards to this, but before we begin, let us recall two important elements put forward by survey methodology, as they are able to influence our perspective.

The first one is the idea of Total Survey Error (TSE, see for example Weisberg, 2005), which argues that researchers should consider all the elements affecting the quality of a survey and take into account the constraints. In this sense, weighting is certainly one parameter that influences the quality of the survey. Some of our colleagues have even proposed making it a quality criteria from a comparative perspective (Zieliński, Powałko, & Kołczyńska, 2018), even if they insist on formal aspects rather than on research-oriented definitions of quality. At the same time, weighting is influenced by elements of design and implementation. In this context, the idea that weighting is a tool for non-response correction is enough to convince researchers that weighting, either with a single weight or a set of weights, could be seen as crucial when considering surveys, but not as an isolated topic, independent of the whole survey's design.

The comparative perspective is the second crucial aspect. An oft asked question is whether the best comparability is obtained using an identical procedure in each component of the comparative survey, most often in each country, or if it is better to adapt the procedure in order to find an optimal strategy in each case (Lynn, 2003). This is clearly a challenge when considering weighting.

These two points, TSE and comparison, were explicitly taken forward when we were drafting the *Sage Handbook of Survey Methodology*, which contains chapters dedicated to sampling and weighting (Gabler & Häder, 2016; Lavallée & Beaumont, 2016, Tillé & Matei, 2016). The later *Palgrave Handbook* by Valette & Krosnick (2018) also posed the question of weighting, such as in DeBell (2018a & 2018b), in contrast to other handbooks, such as the one written by de Leeuw et al. (2008). In summary, the attention given to weighting is quite different according to the textbooks considered. Let us take a closer look at two classical ones.

Good survey methodology books, such as the one by Groves et al. (2004), mention weighting as a standard procedure, with a division between weights linked to the design and those correcting for some events as non-response or other post-stratification adjustments.[1] The *Guidelines for Best Practice in Cross-Cultural Surveys* takes the same approach:

> "To help correct for these differences, sampling statisticians create weights to reduce the sampling bias of the estimates and to compensate for non-coverage and unit non-response. An overall survey weight for each interviewed element typically contains three adjustments: 1) a base weight to adjust for unequal probabilities of selection; 2) an adjustment for sample non-response; and 3) a post-stratification adjustment for the difference between the weighted sample distribution and population distribution on variables that are considered to be related to key outcomes" (Survey Research Center 2016, p. 657).

1 In the literature, the definition of these different steps varies slightly. For simplification, we will insist here on the difference between weights linked to the statistical design and those linked to further corrections, from non-response to calibration.

However, the authors of these guidelines underline from the beginning that such procedures have pros and cons. In short, weighting can reduce coverage bias, non-response bias, and sampling bias at the country or study level. Yet,

> "weighting can increase sampling variance and, when forming non-response adjustment classes, it is assumed that respondents and non-respondents in the same adjustment class are similar: this is a relatively strong assumption. If the accuracy of the official statistics used to create post-stratification adjustments differs by country, comparability across countries can be hampered (Gabler & Häder, 1997). In addition, if the post-stratification adjustments do not dramatically impact the survey estimates, consider not using the adjustment" (Survey Research Center 2016, p. 659-60).

This short summary shows some of the difficulties of using weights in surveys. However, these different points also have to be discussed in more detail. First, we will follow a rather classical line of presentation, discussing design weights and post-stratification weights before proposing general strategies for the analysts.

Design weights

According the literature, there seems to be a consensus on the use of design weights, meaning that the way the survey was designed impacts the quality of data. In the same line, this means that each unit has a known, non-zero probability of being selected. Groves et al. (2004, p. 94) write this: "Probability samples assign each element in the frame a known and non-zero chance to be chosen." In such cases, we can correct the resulting data using these inclusion probabilities. For example, oversampling a region or using a sample based on households is not a problem, as the inclusion probabilities are known, even if the precision of the global sample could be lower according such a design. European Social Survey (ESS), as other surveys, uses the idea of "effective size" and proposes some ways to compute it (Lynn et al., 2007).

In some cases, however, inclusion probabilities are difficult to compute. ESS mentions:

> "The use of random route techniques is strongly discouraged. The reasons for this are a) it is rarely possible to implement such techniques in a way that gives all dwellings even approximately equal selection probabilities; b) it is not possible to accurately estimate these probabilities and therefore to obtain unbiased estimates; and c) the method is easily manipulated by interviewers to their advantage, and in ways that are hard to detect. Instead, as a last resort if no better method is possible, we permit the use of area sampling with field enumeration. How to do this in a way that is consistent with ESS principles is set out in section 2.1.3" (European Social Survey Round 9 Sampling Guidelines: Principles and Implementation The ESS Sampling and Weighting Expert Panel, 26 January 2018, p. 6).[2]

2 A detailed study of what happens in the random route survey in a Western context can be found in Bauer (2014). For a concrete discussion, see Díaz de Rada and Martínez (2014).

Some surveys, such as EU-MIDIS, are using the household size as the basis for design weights, even if all of the elements of the design are not available. We must also mention that design weights based on a variable do not exclude post-stratification weights based on the same variable. Gelman (1997a) mentions, for example, that correcting elements of design based on the size of the household often causes a strong correction because of the differences of lifestyle, and then probability of contact, between single or multiple-person households.

We note that design weights were used by the ESS for many years (Häder & Lynn, 2007) before they turned to adding post-stratification weights in the most recent editions.

What to do?

In summary, we propose first using as much information from the design as possible, even when knowing the difficulty of implementation in some contexts. In the same vein, try to avoid random route and other forms of sampling where inclusion probabilities are difficult or impossible to compute.

Consider design weights, even when considering also using calibration weights in the second step: some authors, such as Haziza and Lesage (2016), argue that a one-step calibration is less robust than a propensity score and then calibration. More research on the links between multi-step weighting, from design to post-stratification, is certainly useful for gaining a better understanding of the impact of weighting.

Post-stratification or calibration

Another family of weights cover corrections based on information of auxiliary variables. We have to consider this a family of methods about which we will not go into technical detail.[3]

The condition to use auxiliary variables is, of course, heavily discussed. For example Särndall and Lundström (2005) provide a literature review on this (p. 129 et seq.). In the same line, Poh and Scheuren (1983) insist on many issues in the field. Among these issues is the tendency to choose weightings for convenience rather than appropriateness on one side, or the importance for secondary analysis to provide a set of weights rather than a single variable, and, finally, the importance to keep the impact of non-response to a minimum[4] (p. 180-181), linking theretofore weighting to the survey's full life-cycle. The idea to keep the non-response rate to a minimum in order to reduce the possible bias, and then weighting, is also mentioned by Lynn (1996). The practical result of such a strategy is described in empirical studies, such as the one by van Goor and B. Stuiver (1998).

3 For more details, see Kalton and Flores-Cervantes (2003).

4 In fact, recent literature insists on minimizing non-response bias rather than non-response by itself (Groves, 2006) but this does not change the reasoning presented here.

Traditionally, auxiliary variables and weighting classes were developed based on the availability of variables and the judgment of the statisticians (Ohl & Scheuren, 1983). Predictors of response, key outcome statistics, and domains are considered in this process. Demographic variables, such as age, sex, race, and geography, were — and still are — frequently chosen, even though they may not be effective in reducing bias (Peytcheva & Groves, 2009). Many of these are population-based adjustments that, for the controls, use data from a recent census. Furthermore, when the number of respondents in a cell of the cross-classification of the variables is below the threshold set for the survey, then cells are collapsed to avoid large adjustment factors (Brick, 2013).[5]

In contrast, not using enough categories could also be a problem. For example, in Switzerland, the foreign population is important and is often under-represented in surveys. However, weighting only with the category "foreigner" will give more importance to the foreigners who have answered the survey; these foreigners are perhaps the most integrated and therefore such a weighting schema do not correct the data at all for the category of the less-well-integrated foreigners (Lagana, 2013, Lipps et al., 2011).

Sometimes, some elements of paradata, such as the number of contacts, could also be envisaged as auxiliary variables (Biemer, Cheng, & Wang, 2013). They are interesting variables to consider as being linked to the data-production process. However, the expectations of such a strategy were not always fulfilled. Furthermore, in a comparative frame, where the modes and organisations of the field can differ, they can be a challenge for comparability.

In fact, Little and Vartivarian (2005) and Bethlehem, Cobben, and Schouten (2011) insist that the problem of weighting is the link between response behaviour on one side and the target variable on the other. "The auxiliary variable selection can be summarised" Bethlehem, Cobben, and Schouten (2011) write, "as the search for variables that replace the unknown relation between target variables and response behaviour" (p. 249). In reference to Särndal and Lundström, they continue by stating, "an ideal auxiliary variable has three features: (1) it explains well the response behavior, (2) it explains well the survey variables and (3) it identifies the most important domain for publication of the survey statistics. By important domains, they mean subpopulation that appear in publications of statistics based on survey" (p. 249).

The problem is therefore difficult to solve because the most accessible variables, such as the socio-demographic ones, are not necessarily related to the response behaviour or to the variables of interest. Furthermore, if we are looking to different variables of interest, the set of predictors could be different, meaning that a set of weights will differ from one analysis to another. And, once again, this means that the situation can be different in different countries.

To go further, it may be good to return to the discussion between statisticians. For example, in the discussion of a seminal paper by Gelman (2007a), Lohr (2007) writes,

5 This is frequently the case when considering some minorities under-represented in the main sample. Some astonishing results in a study about culture in Switzerland were linked to a weight of more than four given to a young foreigner with low level of education but very highbrow tastes. The same type of weighting effect was mentioned in the polls for American Presidential Election (Cohn, 2016).

> "Gelman's (2007) paper begins with the statement 'Survey weighting is a mess.' I do not think that survey weighting is a mess, but I do think that many people ask too much of the weights. For any weighting problem, one should begin by defining which of the possibly contradictory goals for the weights are desirable. Social scientists interested primarily in relationships among variables may value optimality under the model above all other features" (p. 175).

Three important points that we have to consider from this text are crucial here: (1) the importance of documentation, which is a challenge, as weights are often quite technical in the way they are built; (2) the distinction between the estimation of a single statistic or the reasoning on models; and (3) that we cannot ask too much of the weights, meaning that we should not invent information that is not present in the data.

In the same line, Little (2012) writes,

> "From a CB [calibrated Bayes] viewpoint, it is useful to distinguish the case where the variables defining the sampling weights (e.g., the strata indicators in Example 1 above) are or are not included as predictors in the model. If they are, then design weighting is unnecessary if the model is correctly specified. However, from a CB perspective, a comparison of estimates from the weighted and unweighted analysis provides an important specification check, since a serious difference between a design-weighted and unweighted estimate is a strong indicator of misspecification of the regression model" (p. 317).

Once again, we have in this quote the idea to consider models and not just designs. How to build the models is, of course, crucial. The stability of the result with or without weights could also indicate the quality of the estimation.[6] All this discussion is often highly technical, even if the consequences can be rather substantial. However, some authors, such as Brick (2013), are insisting that a better knowledge of the crucial argument is often missing:

> "The central problem, in our opinion, is that even after decades of research on nonresponse, we remain woefully ignorant of the causes of nonresponse at a profound level. This may be a harsh critique, given all the progress we have made in many areas. We better understand methods to reduce nonresponse due to noncontact in surveys and have made substantial

6 Speaking about the quality of the weighting procedure, it could happen to obtain negative weights. Valliant et al. (2013, p. 370) mention that these estimators could be unbiased in theory; "However, negative weights could have a serious effects on some domain estimates, and users are generally uncomfortable with weights that are negative. In fact some software packages will not allow negative weights". Bethlehem et al. (2011, p. 265) write, "Negative and large weights are signs of an ill-conceived model for non-response and target variables", or in the same vein, "another reason to have some control over the values of the adjustment weights is that application of linear weighting may produce negative weights. Although theory does not require weights to be positive, negative weights should be avoided, since they are counter-intuitive. Negative weights cause problems in subsequent analysis, and they are an indication that the regression model does not fit the data well" (p. 237).

> strides in this area. We also have a much better understanding of correlates of nonresponse. Over time, studies have replicated the correlations between demographic and geographic variables and nonresponse rates (e.g., Groves and Couper 1998; Stoop et al. 2010). These are important developments but have not led to a profound understanding of the causes of nonresponse" (Brick, 2013, p. 346).

Of course, from a comparative point of view, this is even more difficult as the mechanisms governing participation in the surveys are not only different by survey but also by countries or social groups.

Let us use an example. In a comparative project, imagine taking education as a post-stratification variable. Is it related to the response process? Probably, even if such a relation could be quite different from one country to another. Is it related to the target variables? This is probably true in some cases, but also less true for other analyses. Therefore having such a variable as an explicit control rather than as an obscure weight could allow researchers to learn more about the relations. However, this is not the end of the story. Is it possible to measure education internationally? The response is probably yes, according the work around the International Standard Classification of Education (ISCED). Is it possible, however, to have a common measure of reference in the countries, and on which basis? Even if the European Labour Force Survey (LFS), often used in order to provide data on education in different countries, could be seen as the "best" standard in the field, it is not without criticisms, as the LFS is a survey that is itself adjusted in some ways in the countries. Furthermore, some countries may not be running LFS and, therefore, the question of the homogeneity of sources is a problem that Ortmanns and Schneider (2016) describe clearly.

In summary, in comparative surveys, the question of weighting is complicated not only by the difficulty of having homogeneous and reliable sources for additional information but by the questions of similarities in the response and non-response process, as well as in the link between weighting variables and potentially varying variables of interest. That means that, returning to the alternatives proposed by Lynn (2003), adapting the weighting strategy to the local conditions of each countries could be more adapted than choosing an identical procedure in every country.

What should be done?

According to the arguments discussed so far, we would be careful when calculating and using post-stratification or calibration weights and recall that they can be a set of variables and that they have to be adapted to the research question. When computing point estimates, such as the mean of a variable, a weighting strategy is probably the most appropriate one. However, this is not necessarily the case when looking to models. Regardless, detailed documentation must be available.

This is clearly the strategy proposed by SHARE when they refer to the paper by Solon et al. (2015[2013]) that mentions,

> "In Section II, we distinguished between two types of empirical research: (1) research directed at estimating population descriptive statistics and (2) research directed at estimating causal effects. For the former, weighting is called for when it is needed to make the analysis sample representative of the target population. For the latter, the question of whether and how to weight is more nuanced…. Our overarching recommendation therefore is to take seriously the question in our title: What are we weighting for? Be clear about the reason that you are considering weighted estimation, think carefully about whether the reason really applies, and double-check with appropriate diagnostics" (p. 20).

Population weights

To our knowledge, the literature is not definitive for data users of comparative surveys. In particular, the question of how to take into account the characteristics of each survey in each country is still open -- to begin with, the question of how to take into account differences of size between countries. In a recent publication, Kaminska and Lynn (2017) suggest using routine population weights according to the population of the country. This proposition is also interesting to mention because Lynn more readily defended the sole use of design weights in the first edition of the ESS.

Going back in the history of statistics, Kish (1999[2003]) discusses six options when considering "multiple population" surveys, which is the case of international projects:

1. Do not combine the data from different countries. This means to exclude comparisons.
2. Harmonize survey measurements in each case, but do not use a common analysis. This just produces statistics by country and compares them.
3. Use an equal weight for each country in the combined analysis.
4. Weight with the sample size, and eventually with the effective size of each sample, knowing that there are more institutional reasons than statistical ones governing the size of each national sample.
5. Population weight should pay careful attention to the reference population (inhabitants, citizens, voters, etc.) according the analysis envisaged.
6. "Post-stratification weights" are where strata could be constituted from one country and another strata from many countries. If we follow such logic in some international comparative projects, each continent or "big region" could be considered strata. Of course, the theoretical rationale for such a construction has to be established.

Such a choice must be made according substantive reasons. Otherwise, Kish's (1999, p. 295 [in the 2003 reprint]) personal preference is for solutions 4 or 5 rather than solutions 1 or 2, meaning researchers should take into account the comparative dimension. He mentions that with some giants like China or India, solution 3 or 5 could be difficult; 5 because of the increase in variance if the sizes are too different. In the case of ISSP, for example, Iceland has 360,000 inhabitants and China 1,386,000,000 nearly 3,850 times more than that of Iceland, which means there is probably also much more variance in many of the indicators.

But according Kish, it is also important to take into account the impact on bias. In other words, if the country has no influence, then the impact of such weights will be negligible. The last recommendation given by Kish is "that the combination of population surveys into multi-population statistics needs a good deal of research, both empirical and theoretical – and especially together" (Kish, 1999[2003], p. 295).

What should be done?

In the absence of the researches proposed by Kish, we suggest a differentiated strategy:

1. If the goal is to produce a descriptive statistic for all countries included, then follow Kaminska and Lynn's (2017) work and use population weights for the total sample.
2. If the idea is to compare the situations in different countries from an aggregate perspective, then there is no need to use such population weights, as the figures are produced country by country. This is the case in Welzel and Inglehart's (2016) approach, except when the values are the mean (by country) of a factor analysis of the pooled data set, where the question of weighting according to population could be asked again.
3. If scalar equivalence through multi-group factor analysis is confirmed, then the country alone does not play any role. However, this does not mean that any comparison is impossible if this condition is not fulfilled.
4. Otherwise, a multilevel perspective takes into account the country's effect, and then we do not need population weights. This does not mean that weighting is not necessary in a multilevel perspective. As Pfeffermann et al. (1998) note, multilevel does not mean 'do not weight:' "When the sample selection probabilities are related to the response variable even after conditioning on covariates of interest, the conventional estimators of the model parameters may be (asymptotically) biased" (p. 24).

The category "country" plays an ambiguous role in the context of international comparative surveys. On one side, the field is organised most often along this division, which means that we have a lot of effects that are linked to field organisation, "survey climate" (Loosveldt & Joye, 2016) and, sometimes, modes. On the other side, by having specific institutions, policies, and so on, "country" is an important, aggregated category and has to be taken into account in the debate on individualistic-ecological fallacies (Alker, 1969). But, when taking context into consideration, other aggregates, such as regions or even social groups, to begin with social classes or networks, have to be considered in the modelling strategy.

Conclusion

In the end, we can propose some recommendations for producers and users of comparative surveys. For survey producers, the first point is certainly documentation. Documentation is even more important from an ethical perspective; some authors see weighting as a form of data manipulation. Of course, we know the difficulty of presenting such technical information, even more when considering the extraordinary amount of documentation that has to be provided, but this challenge is worth the effort. This challenge is also underlined, for example, by DeBell (2018a, p. 161) asking about weights: "A) More accessible guidance; B) Increased transparency of methods; C) Clearly replicable methods; D) Standardized/comparable methods rather than ad hoc approaches." We add to this the proposal to consider different sets of weights, according to approaches and analyses. Furthermore, from a comparative perspective, some thoughts about the processes that produce the need for weighting and a discussion of their similarities or dissimilarities between countries could be useful.

For survey users, more attention should also be given to the question of weights. In particular, a challenge is how to consider them when modelling. In this sense, the theoretical perspective on what we are doing in analysis is crucial. In the same line, how to consider the variable "country" is also crucial. We would also advocate for more exchanges between statisticians and social scientists to better understand the underlying conceptualisation of each discipline. Furthermore, we sometimes have the impression that survey users are more oriented towards analysis than towards understanding data quality and limitations. The lack of attention given to weights could be related to the perception that it is a problem for survey producers and not survey users. If we are correct, the issue of weighting could also be more incorporated in the regular training and in summer schools but this is a question to discuss explicitly.

In any case, we need more analysis on the effect of weighting and the way to take such a factor into account, specifically in regards to the idea of combining theoretical and empirical perspectives. Kish (1994) made already a similar proposal twenty-five years ago and this is still relevant today. We hope that this contribution is a call to "open the black box" and experiment further in the field, specifically from a more comparative perspective.

Dominique Joye is professor of sociology (emeritus) at University of Lausanne (and affiliated researcher at FORS, forscenter.ch)

Marlène Sapin is Senior Researcher at the Swiss Center of Expertise in Social Sciences FORS and in the NCCR LIVES – Overcoming Vulnerability: Life Course Perspectives, at the University of Lausanne.

Christof Wolf is President of GESIS Leibniz-Institute for Social Sciences and Professor of Sociology at University Mannheim

References

Alker, H. R. (1969). "A typology of ecological fallacies." In H. Dogan & S. Rokkan (Eds.), *Quantitative ecological analysis in the social sciences* (pp. 69-86). MIT press, London.

Bauer J.J. (2014) "Selection Errors of Random Route Samples." *Sociological Methods & Research*, Vol. 43(3) 519-544.

Bethlehem J. (2008) "Weighting" in Lavrakas P.J., *Encyclopedia of Survey Research Methods*, Vol. 2, pp. 957-960.

Bethlehem J. (2009) *Applied Survey Methods*, Wiley, Hoboken.

Bethlehem J., Cobben F. & Schouten B. (2011) *Handbook of Nonresponse in Household Surveys*, Wiley, Hoboken N.J.

Biemer P.P., Chen P. & Wang K. (2013) "Using level of effort paradata in non-response adjustments with application to field surveys" *Journal of the Royal Statistical Society, Statistics in Society, Series A*, Vol 176 (1), pp. 147-168.

Brick J.M. (2013) "Unit Nonresponse and Weighting Adjustments: A Critical Review" *Journal of Official Statistics*, Vol. 29, No. 3, pp. 329–353.

Cohn N. (2016) "How One 19-Year Old Illinois Man is Distorting National Polling Averages", *New York Time*, Oct. 12, 2016, Retrieved from https://www.nytimes.com/2016/10/13/upshot/how-one-19-year-old-illinois-man-is-distorting-national-polling-averages.html the 30th of October 2019.

Chatrchi, G., Duval, M. C., Brisebois, F., & Thomas, S. (2015) "The Impact of Typical Survey Weighting Adjustments on the Design Effect: A Case Study" *Survey Methods: Insights from the Field: Practical Issues and 'How to' Approach*. Retrieved from https://surveyinsights.org/?p=4919

DeBell M. (2018a) "Best Practices for Creating Survey Weights" inVannette D.L. & Krosnick J.A. (eds.), *The Palgrave Handbook of Survey Research*, Palgrave, pp. 159-164.

DeBell M. (2018B) "Computation of Survey Weights" in Vannette D.L. & Krosnick J.A. (eds.), *The Palgrave Handbook of Survey Research*, Palgrave, pp. 519-527.

Díaz de Rada V. & Martínez Martín V. (2014) "Random Route and Quota Sampling: Do They Offer Any Advantage over Probably Sampling Methods?" *Open Journal of Statistics*, 2014, 4, 391-401.

European Union Agency for Fundamental Rights (2009) *EU-MIDIS Technical Report, Methodology, Sampling and Fieldwork*, retrieved from http://fra.europa.eu/eu-midis, accessed on 23rd of October 2019.

Gabler, S., & Häder, S. (1997) "Deviations from the population and optimal weights" in W. E. Saris, & M. Kaase (eds) *Eurobarometer: measurement instruments for opinions in Europe* (ZUMA-Nachrichten Spezial 2) (pp. 32-44). Mannheim, Germany: ZUMA.

Gabler S. & Häder S. (2016) "Special Challenges of Sampling for Comparative Surveys", in Wolf C., Joye D., Smith T.W. & Fu Y. (eds) *The Sage Handbook of Survey Methodology*, Sage, London, pp. 346-355.

Gelman A. & Carlin J.B. (2002) "Poststratification and Weighting Adjustment" in Groves R.M. et al., *Survey Nonresponse*, Wiley, Hoboken.

Gelman A. (2007a) "Struggles with Survey Weighting and Regression Modeling" in *Statistical Science*, Vol. 22, No. 2, pp. 153–164.

Gelman A. (2007b) "Rejoinder: Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, Vol. 22, No. 2, pp. 184-188.

van Goor H. & Stuiver B. (1998) "Can Weighting Compensate for Nonresponse Bias in a Dependent Variable? An Evaluation of Weighting Methods to Correct for Substantive Bias in a Mail Survey among Dutch Municipalities" *Social Science Research*, 27, 481–499.

Groves R.M. (2006) "Nonresponse Rates and Nonresponse Bias in Household Surveys" *Public Opinion Quarterly*, Vol. 70, No. 5, pp. 646-675.

Groves R.M. Floyd J. F. Jr., Couper M. P., Lepkowski J. M., Singer E., and Tourangeau R. (2004) *Survey Methodology*, Wiley, Hoboken N.J.

Häder S. & Lynn P. (2007) "How representative can a multi-nation survey be?" in Jowell R. et al., *Measuring Attitudes Cross-Nationally*, Sage London.

Haziza, D. & Lesage, E. (2016). "A discussion of weighting procedures for unit nonresponse" *Journal of Official Statistics*, 32, 129-145.

Joye D. Wolf C, Smith T.W. & Fu Y. (2016) "Survey Methodology, Challenges and Principles" in Wolf C., Joye D., Smith T.W. & Fu Y. (eds) *The Sage Handbook of Survey Methodology*, Sage, London, pp. 5-16.

Kalton G. (2002) "Models in the Practice of Survey Sampling (Revisited)" *Journal of Official Statistics*, Vol. 18, No. 2, pp. 129 -54.

Kalton G. & Flores-Cervantes I. (2003) "Weighting Methods" *Journal of Social Statistics*, Vol. 19, No. 2, pp. 81-97.

Kaminska O. & Lynn P. (2017) "Survey-based cross-country comparisons where countries vary in sample design: issues and solutions" *Journal of Official Statistics*, Vol. 33, No. 1, pp. 123–136.

Kish L. (1994) "Multipopulation Survey Designs: Five Types with Seven Shared Aspects", *International Statistical Review / Revue Internationale de Statistique*, Vol. 62, No. 2, pp. 167-186.

Kish L. (1999) "Cumulating/Combining Population Surveys" *Survey Methodology*, 15, pp. 129-138, reprinted in Kalton G. & Heeringa S. (eds) (2003) *Leslie Kish, selected Papers*, Wiley, Hoboken.

Laganà F. et al. (2013) "National minorities and their representation in social surveys: which practices make a difference?" *Quality & Quantity* (2013) 47:1287–1314.

Lavallée, P. & Beaumont, J. F. (2015) "Why We Should Put Some Weight on Weights" *Survey Insights: Methods from the Field*, Weighting: Practical Issues and 'How to' Approach, Invited article, Retrieved from http://surveyinsights.org/?p=6255, accessed on 23rd of October 2019.

Lavallée, P. & Beaumont, J. F. (2016) "Weighting: Principles and Practicalities", in Wolf C., Joye D., Smith T.W. & Fu Y. (eds) *The Sage Handbook of Survey Methodology*, Sage, London, pp. 460-476.

de Leeuw E.D., Hox J.J., Dillman D.A. (2008) *International Handbook of Survey Methodology*, Routledge.

Lipps, O., Laganà, F., Pollien, A., & Gianettoni, L. (2011). National minorities and their representation in Swiss surveys (I): Providing evidence and analysing causes for their under-representation. in Lago M. & Font J. (eds.), *Surveying ethnic minorities and immigrant populations: methodological challenges and research strategies*. Amsterdam University Press.

Little R.J: (2007) "Comment: Struggles with Survey Weighting and Regression Modeling," *Statistical Science*, Vol. 22, No. 2, pp. 171-174.

Little R.J. (2012) "Calibrated Bayes: an Alternative Inferential Paradigm for Official Statistics" *Journal of Official Statistics*, 28, 3, 309-372.

Little R.J. & Vartivarian S. (2005) Does Weighting for Nonresponse Increase the Variance of Survey Means?" *Survey Methodology*, Vol. 31, No. 2, pp. 161-168.

Lohr S.L. (2007) "Comment: Struggles with Survey Weighting and Regression Modelling", *Statistical Science*, Vol. 22, No. 2, pp. 175-178.

Lynn P. (1996) "Weighting for nonresponse" in *Survey and statistical computing 1996: proceedings of the second ASC international conference,* Imperial College, London, UK, September 11-13 1996, retrieved from https://pdfs.semanticscholar.org/1e7d/d794cbaf774ecfe578b493859f29ea6c13f2.pdf the 30th of October 2019

Lynn P. (2003) "Developing quality standards for cross-national survey research: five approaches", I*nt. J. Social Research Methodology*, 6:4, 323-336.

Lynn P., Häder S. & Gabler S. (2007) "Methods for Achieving Equivalence of Samples in Cross-National Surveys: The European Social Survey Experience" *Journal of Official Statistics*, Vol. 23, No. 1, pp. 107–124.

Loosveldt, G., & Joye, D.. (2016). "Defining and assessing survey climate".in Wolf C, Joye, D., Smith, T., & Fu, Y. (eds), *The SAGE Handbook of Survey Methodology* SAGE, London, pp. 67-76.

Oh H.L. & Scheuren F.J. (1983) "Weighting Adjustement for Unit Nonresponse" in Madow W.G., Olkin I. & Rubin D. B. (eds) *Incomplete Data in Sample Surveys* Vol. 2, Part 4, pp. 143-184.

Ortmanns V. & Schneider S.L. (2016) "Can we assess representativeness of cross-national surveys using the education variable?" *Survey Research Methods* Vol. 10, No. 3, pp. 189-210.

Peytcheva E & Groves R.M. (2009) "Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates" *Journal of Official Statistics*, Vol. 25, No. 2, 2009, pp. 193–201.

Pfeffermann D. (1993) The Role of Sampling Weights When Modeling Survey Data, *International Statistical Review / Revue Internationale de Statistique*, Vol. 61, No. 2, pp. 317-337.

Pfeffermann D.; Skinner C. J.; HolmesD. J.; Goldstein H.; Rasbash J. (1998) "Weighting for Unequal Selection Probabilities in Multilevel Models", *Journal of the Royal Statistical Society*. Series B (Statistical Methodology), Vol. 60, No. 1, pp. 23-40.

Pfefffermann D. (2011) Modelling of complex survey data: Why model? Why is it a problem? How can we approach it? *Survey Methodology*, Vol. 37, No. 2, pp. 115-136.

Potter F. & Zeng Y (2015) "Methods and Issues in Trimming Extreme Weights in Sample Surveys" presentation in *Proceedings of the Joint Statistical Meetings 2015 Survey Research Methods Section*, Washington, available at http://www.asasrms.org/Proceedings/y2015/files/234115.pdf, accessed on 30 of October 2019

Särsndal C.-E. & Lundstöm S. (2005) *Estimation in Surveys with Nonresponse*, Wiley, Hoboken.

Solon G. & Haider S.J. & Wooldridge, J.M. (2015) "What Are We Weighting For?" *Journal of Human Resources*, University of Wisconsin Press, vol. 50(2), pages 301-316. Also published in 2013 as working paper at the address https://www.nber.org/papers/w18859.pdf, accessed 31st of October 2019

Survey Research Center. (2016). *Guidelines for Best Practice in Cross-Cultural Surveys*. Ann Arbor, MI: Survey Research Center, Institute for Social Research, University of Michigan. http://www.ccsg.isr.umich.edu/, accessed 31th of October 2019

Tillé Y & Matei A. (2016) "Basis of Sampling for Survey Research" in Wolf C., Joye D., Smith T.W. & Fu Y. (eds) *The Sage Handbook of Survey Methodology*, Sage, London, pp. 311-328.

Vannette D.L. & Krosnick J.A. (eds.) *The Palgrave Handbook of Survey Research*, Palgrave.

Valliant R., Dever J.A., & Kreuter F. (2013) *Practical Tools for Designing and Weighting Survey Samples*, Springer, New York, Heidelberg, Dordrecht & London.

Weisberg H.F. (2005) *The Total Survey Error Approach: A Guide to the New Science of Survey Research*, University of Chicago Press, Chicago.

Welzel C & Inglehart R.F. (2016) "Misconceptions of Measurement Equivalence: Time for a Paradigm Shift", *Comparative Political Studies*, Vol. 49(8), pp. 1068–1094.

Wolf C., Joye D., Smith T.W. & Fu Y. (eds) (2016) *The Sage Handbook of Survey Methodology*, Sage, London.

Zieliński M.W, Powałko P. & Kołczyńska M. (2019) "The Past, Present, and Future of Statistical Weights in International Survey Projects: Implications for Survey Data Harmonization" in Johnson T. P., Pennel B.-E., Stoop I. & Dorer B. (eds) *Advances in comparative survey methods. Multinational, Multiregional, and Multicultural Contexts (3MC)*, Wiley, Hoboken.

Combining Cognitive Interviews and Web Probing for Cross-Cultural Cognitive Pretesting: The Case of the European Working Conditions Surveys

by Patricia Hadler and Agnès Parent-Thirion

This note presents a multi-method cognitive pretest that was coordinated and consolidated by the GESIS Pretest Lab and commissioned by EUROFOUND for the 6th wave of the European Working Conditions Surveys (EWCS) (Eurofound 2017). The aim of this note is to exemplify how the pretesting methods of face-to-face cognitive interviewing and web probing can be combined to rise to the challenges of cross-cultural cognitive pretesting.

The European Working Conditions Surveys – Aims and Challenges

The EWCS examines people's working lives and the quality of their jobs in order to provide comparable information on working conditions across European countries. Collecting comparable data for EWCS has become more challenging. The first challenge is to ensure cross country comparability across an increasing number of participating countries. Twelve countries participated in the survey's initial wave in 1990; the upcoming Wave 7 will comprise 37 countries. The second challenge is to collect comparable data across heterogeneous working populations. The fragmentation of work and employment has resulted with more people in "atypical" working situations, such as multiple jobs, combinations of employment and self-employment, civil employment contracts, and workers on the margins.

EUROFOUND commissioned the GESIS Pretest Lab to carry out cognitive pretesting of the EWCS questionnaire from Wave 6. There were two aims: The first aim was to examine how well the questions that assess employment status apply to atypical groups of workers. The second aim lay in comparing how employees, as compared to self-employed respondents, understood items that pertain to job quality.

The Rationale of Combining Web Probing and Cognitive Interviews

Cognitive interviewing has proven to be a very useful tool for the development of comparative questionnaires across countries and groups (cf. Miller et al., 2014). The unmatched strength of face-to-face cognitive interviews is to provide in-depth insights on respondents' cognitive response processes (Beatty & Willis, 2007). While interviewers generally use a standardized interview protocol, they can use emergent probing follow up on complex issues. Although carrying out cognitive interviews in multiple countries is recommended (Willis, 2015; Willis & Miller, 2011), it is both expensive and time-consuming.

For this reason, in the context of cross-cultural surveys, web probing has become an increasingly popular pretesting method (Behr, Braun, Kaczmirek & Bandilla, 2014). The lack of interviewer and organization effects facilitates a similar approach across countries and languages, as well as a cost-efficient and quick method of data collection. In terms of cognitive testing, it is particularly useful to gather information on the comprehension of single terms and the reasons for selecting answer options (Behr, Meitinger, Braun & Kaczmirek, 2017).

In order to gain deep insights into individual working situations, and also collect data from several countries and diverse employment situations, we created a research design that combined web probing and cognitive interviewing as both methods have complementary strengths and weaknesses (for comparisons of the methods, see Lenzner & Neuert, 2017; Meitinger & Behr, 2016).

We decided to first examine the job quality indices using the method of web probing. Participating in the cognitive online pretest was 365 respondents in the UK, Germany, and Poland. The (comparably) high case number not only allowed us to compare employees and the self-employed, but even to distinguish between self-employed with and without employees of their own.

Following the web probing study, cognitive interviews were carried out in Germany and Poland by GESIS and Kantar. These interviews were used to collect insights on atypical working populations (which would have been difficult to recruit in an online access panel), and gain deeper insights into respondents' cognitive process of survey response.

Pretest Results and Implementation

To demonstrate the complementary benefits, we provide two examples of items that were tested using both methods (see Hadler, Neuert, Lenzner, & Menold, 2018).

One item from the job quality index on "Social Environment" is: "*Your colleagues help and support you*" (Question 61a). It is answered using a full-labeled five-point scale ranging from "always" over "most of the time," "sometimes," "rarely," and "never." If the respondent refuses to answer the question, the interviewer records non-response; however, this response option is not visible for the respondent.

During web probing, we deviated from the interviewer-assisted procedure and explicitly offered the non-response option "not applicable" to respondents. Among self-employed

respondents, 55 percent (UK), 48 percent (Germany), and 33 percent (Poland) chose the option "not applicable"; of respondents who are employees, only between three percent and five percent chose this option. The subsequent probing questions established that all respondents who chose "not applicable" had no colleagues to whom they could refer this statement. However, of the eight respondents who claimed that they "never" received help or support from colleagues, five also stated as a reason that they have no colleagues; the same applies to 11 of 20 respondents who answered "rarely."

The subsequent cognitive interviews simulated the actual survey situation by only recording non-response when the respondent did not offer an answer. Of the eight participants who were self-employed, only two Polish participants chose not to answer the item. Several participants understood the term "colleague" to include other self-employed people working in the same field. Several self-employed participants referred to their employees while answering this question.

Based on these findings, the questionnaire design team at EUROFOUND decided to establish two versions of the item battery on social environment at work: one for employees ("your colleagues help and support you") and one for the self-employed ("your colleagues or peers help and support you"). Another item from the job quality index on "Working Time" asks "*Over the last 12 months, how often have you worked in your free time to meet work demands?***"** (Question 46). Five response options cover a range from "daily" over "several times a week," "several times a month," "less often," and "never."

Web probing was used to establish how well self-employed respondents can distinguish between work and free time. Self-employed respondents were more likely to experience difficulties in distinguishing between work and free time than employed respondents. However, this did not apply to all self-employed respondents; those who had clearly defined and regular working hours – for instance because they own a store – experienced no issues. No notable differences in question comprehension between countries were found during web probing.

The cognitive interviews revealed that 11 of 16 German participants spontaneously commented on the question wording. All remarks pertained to the German words for "free time" and "work demands," especially trying to decide whether working in one's free time is supposed to mean "working overtime." This uncertainty did not arise with any Polish participants. Also, cognitive interviewing revealed that, across all types of workers, a wide range of work-related activities that are done outside of official working hours are inconsistently included and excluded as "working in one's free time."

The questionnaire design team recognized that the issue of distinguishing both working time and work-related activities from free time was crucial to the quality of the survey data and affects several survey questions. The next wave will specifically collect information on work-related communication in one's free time because pretesting uncovered this as one of the main work-related free time activities. Instructions for the existing question were added for self-employed respondents. Also, both contractual time and actual working time will be collected to better quantify unpaid overtime.

The need to ask clear, unambiguous survey questions that apply to -- and can thus be correctly answered by -- workers in a variety of employment situations and across multiple countries was acknowledged, and numerous revisions are being implemented into the next wave of the European Working Conditions Surveys. Next to the few examples in this report, this has led to a stronger use of filters and population-adapted wording to make questions equally relevant and comprehensible to all respondents. This will contribute to the production of valid measurements and facilitate harmonization of measures across different groups of workers and countries.

Outlook: Cross-Cultural Multi-Method Cognitive Pretesting and Harmonization

Combining web probing and cognitive interviewing proved an effective way of assessing and improving equivalent measures. Cognitive interviews provided in-depth insights for atypical groups of respondents, while the use of web probing made it possible to compare subgroups of respondents, especially across countries. We fielded the web probing study first, and used subsequent cognitive interviews to follow up on remaining questions. However, depending on the research questions, in other cases, beginning with cognitive interviews may be better suited to the purposes of the study. Future research will hopefully contribute to the establishment of best practices in multi-mode and cross-cultural cognitive pretesting to meet the challenges of harmonization in cross-national surveys.

NB: A detailed project report can be found in the pretest database of the GESIS Pretest Lab: https://pretest.gesis.org/pretestProjekt/Eurofound-Parent-Thirion-Preparation-of-the-7th-European-Working-Conditions-Survey-%28EWCS%29-Post-test-of-the-6th-EWCS

Information on the services of the GESIS pretest lab can be found on the website of GESIS https://www.gesis.org/en/services/study-planning/cognitive-pretesting

Information on the European Working Conditions Surveys (EWCS) can be found on the website of Eurofound: http://eurofound.europa.eu/surveys/european-working-conditions-surveys

Patricia Hadler is a researcher and doctoral student at GESIS – Leibniz Institute for the Social Sciences.

Agnès Parent-Thirion is Senior Research Manager at EUROFOUND.

References

Beatty, Paul; & Willis, Gordon (2007). Research Synthesis. The Practice of Cognitive Interviewing. *Public Opinion Quarterly, 71*(2), 287–311.

Behr, Dorothée; Braun, Michael; Kaczmirek, Lars; Bandilla, Wolfgang (2014). Item comparability in cross-national surveys. Results from asking probing questions in cross-national web surveys about attitudes towards civil disobedience. *Quality & Quantity, 48*(1), 127-148.

Eurofound (2017). *Sixth European Working Conditions Survey – Overview report (2017 update)*, Publications Office of the European Union, Luxembourg. Retrieved from https://www.eurofound.europa.eu/sites/default/files/ef_publication/field_ef_document/ef1634en.pdf

Behr, Dorothée; Meitinger, Katharina; Braun, Michael; Kaczmirek, Lars (2017). Web probing – implementing probing techniques from cognitive interviewing in web surveys with the goal to assess the validity of survey questions. Mannheim, GESIS – Leibniz Institute for the Social Sciences (GESIS - Survey Guidelines).

Hadler, P., Neuert, C., Lenzner, T. & Menold, N. (2018): Preparation of the 7th European Working Conditions Survey (EWCS) – Post test of the 6th EWCS. Final Report April-November 2018. GESIS Project report. Version: 1.0. GESIS - Pretestlabor.

Lenzner, Timo; & Neuert, Cornelia (2017). Pretesting Survey Questions Via Web Probing – Does it Produce Similar Results to Face-to-Face Cognitive Interviewing? *Survey Practice, 10*(4), 1-11.

Meitinger, Katharina; & Behr, Dorothée (2016). Comparing Cognitive Interviewing and Online Probing: Do They Find Similar Results? *Field Methods, 28*(4), 363–380.

Miller, Kristen; Chepp, Valerie; Willson, Stephanie; Padilla, Jose-Luis (Eds.) (2014). *Cognitive Interviewing Methodology*. Wiley Series in Survey Methodology. New York: Wiley.

Willis, Gordon B. (2015). Research Synthesis. The Practice of Cross-Cultural Cognitive Interviewing. *Public Opinion Quarterly, 79*(S1), 359–395.

Willis, Gordon; & Miller, Kristin (2011). Cross-Cultural Cognitive Interviewing: Seeking Comparability and Enhancing Understanding. *Field Methods, 23*(4), 331–341.

Do Differences in Fieldwork Period Affect Survey Comparability? Examining World Values Survey and European Values Study in Romania, 2017 – 2018

By Bogdan Voicu

Comparability across surveys depends on using the same measurement in terms of scales and conditions (Harkness, 1999). Means for testing measurement invariance were developed to check comparability of the measurement tool across cultures and groups (Davidov, 2010; Kankaraš et al, 2010). However, even when the tool is reliable, measurements might not be comparable if the conditions under which data are collected are not similar enough.

This research-note examines the extent to which the fieldwork period, i.e. time of data collection, is associated with the quality of the estimates. Using pretty similar surveys, I focus on

four different dependent variables that differ in terms of scale, and come from four different fields, which means that they have attached different assumptions related to influence of the period of the year when data were collected. The measurement specifications are of secondary importance to this research-note; therefore, I introduce first the type of scales, then I discuss the time-related assumptions, and only in the end I explain the exact measurement. The focus is not on what the variables measure, but on how time may influence the estimates. The four measures are predicted with basic socio-demographic indicators, along with indicators of when data were collected, and a control for the survey from which they are extracted.

To put it simply, I test whether one gets the same estimates on the same population from which different samples were extracted using identical survey methods. What varies is the questionnaire and the date of data collection. The questionnaires differ slightly, they include the same topic, and are of same length, meaning the time allotted to complete them are relatively similar. In other words, potential influence from uncontrolled confounders is very little, if any.

I use the 2017/2018 Romanian waves of World Values Survey (WVS) and European Values Study (EVS). WVS was collected from November 2017 to April 2018. EVS fieldwork took place from February 2018 to May 2018. There is little overlap between the two surveys, since most of the WVS questionnaires were filled in before EVS started, as indicated in Figure 1.

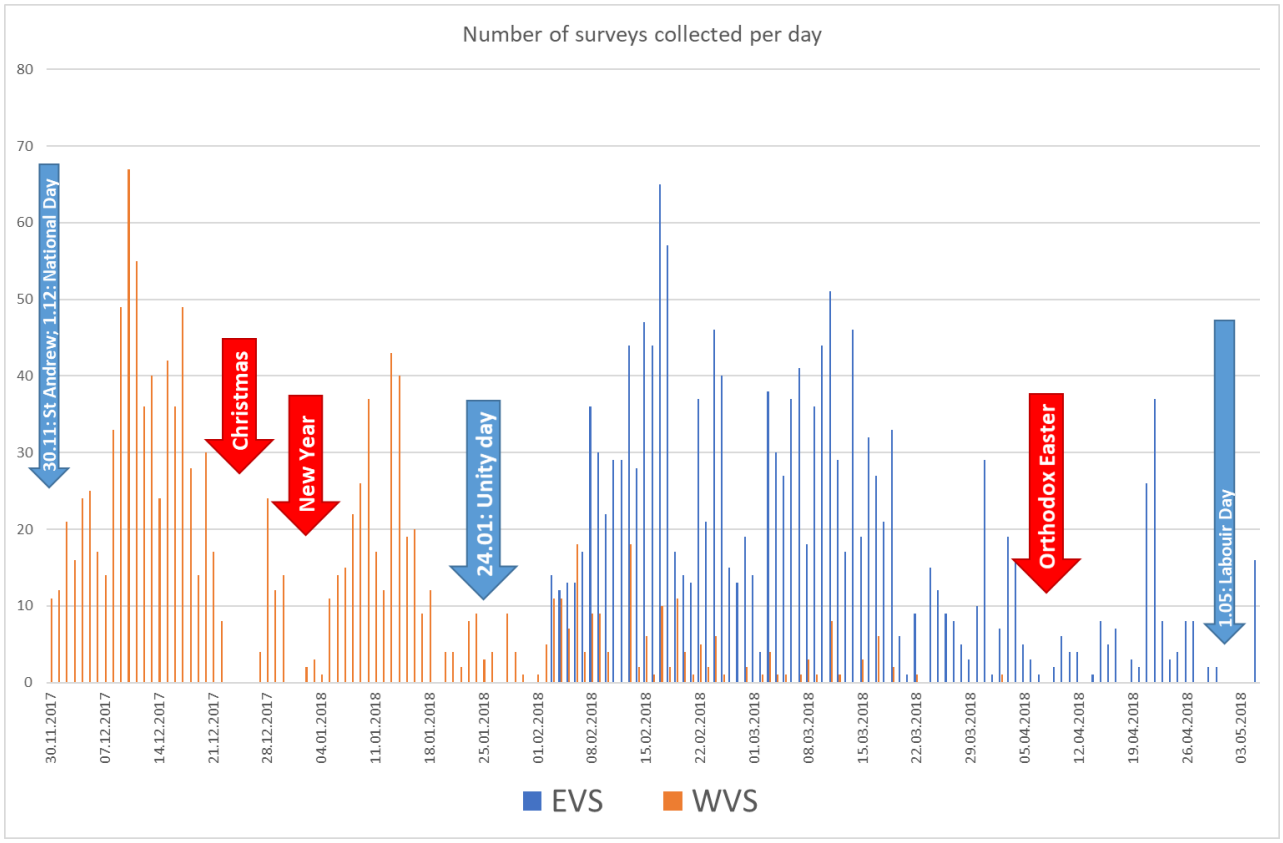


Figure 1. Timing of data collection for EVS and WVS Romania 2017/2018, and public holidays

Fieldwork was carried out by the same consortium of two data collecting agencies (IRES and MMT). The research design was provided by the central teams of EVS and WVS, which cooperated in producing similar questionnaires (roughly three quarters of the questionnaires are identical). The Romanian Group for Studying Social Values adapted everything to Romanian conditions, prepared the translation of the questionnaires, and included almost the same country-specific items in the two surveys. The sampling plan followed identical specifications for both EVS and WVS and was done by the same person (professor Mircea Comşa, part of the Romanian Group for Studying Social Values). The fieldwork interviewers were trained in the same manner, and many worked within the two data collection activities. Finally, the weighting systems for the two samples are done in the same way, such as to make the sample structure similar to the same pattern in the general adult population.

The main difference between the two surveys is the period of the year in which these data were collected. This means that almost everything is the same, and one may use the opportunity to check the extent to which time intervenes as a distortion between the two series of estimates. Between November 2017 and May 2018 the most salient societal events were public holidays, including Easter and Christmas. No major political event took place.[1]

As mentioned, I study cross-survey differences with respect to four dependent variables. They reflect four different ways to measure concepts, and include: (1) a single-item measurement that is categorical; (2) a single-item measure that is continuous; (3) a count indicator, and (4) one that is also additive, but the components are weighted before aggregation.

The four measures come with different assumptions related to whether fieldwork period may influence respondents' answers. The first measures religious belief and is expected to vary more than the others across time. Salient events such as Christmas and Easter mark the dynamics of daily life and may temporarily change religiousness. The second variable is life satisfaction, which is supposed to be less sensitive to personal and societal events (Ehrhardt et al, 2000), but it was shown, in some cases, to depend on societal-level emotional situations, such as elections (Voicu, 2012). The third measure is postmaterialism, which is supposed to vary less with time-related personal events. The fourth measure is trust in political institutions, which is likely to vary across time (Tufiş, 2012).

Religious belief is measured through a package of four items, indicating whether the respondent beliefs in hell, god, heaven, and life after death. Each variable accepts dichotomous answers (yes/no). I have computed a very simple indicator counting how many of these four beliefs were checked by the respondent. The new belief indicator ranges from 0 to 4.

Life satisfaction is a single-item measurement, ranging from 1 to 10. The answer "don't know" (DK) and the refuse to answer (NA) were treated as missing cases (they count for roughly four percent out of the total sample and are too few to properly test whether they vary depending on time of data collection).

[1] The major political protests in Romania that garnered worldwide headlines occurred in January and February 2017, many months before WVS fieldwork started; and in August 2018, after WVS and EVS fieldwork.

For postmaterialism, I use the first choice within the classic Inglehart measure (Inglehart, Abramson 1999), namely mentioning the top priority out of four: (i) Maintaining order, (ii) Giving people more say, (iii) Fighting rising prices, (iv) Protecting freedom of speech. Again, DK and NA (two percent of the sample) were excluded from analysis. The typical analysis of the item supposes giving one point for each choice of the second and the fourth alternatives and subtracting one point for each choice of the remaining ones. To increase comparability across surveys, I use the percentage of those who chose the first alternative, as mentioned.

Finally, for trust in political institutions, three 4-points items are used, denoting trust in Parliament, Government, and Political Parties. A factor score was computed based on the adequacy of factor analysis: KMO=0.738, factor loadings are around 0.8 (estimated with maximum likelihood). Table 1 displays the 95% confidence intervals for the continuous dependent variables. One may easily observe that Religious belief and Political trust have similar estimated means for Romanian WVS 2017 and EVS 2018, given the overlapping of the 95% CIs. Life Satisfaction has significantly higher scores in the WVS sample, that is, during the winter break, as compared to the springtime collected EVS sample.

Dependent variable	N	**Range**		**95% CI for mean: WVS**		**95% CI for mean: EVS**	
		Min	Max	Lower bound	Upper bound	Lower bound	Upper bound
Religious belief	2871	0	4	2.52	2.69	2.56	2.70
Life Satisfaction	2753	1	10	3.96	4.26	3.67	3.94
Postmaterialism	2802	four categories*		categorical*			
Political trust	2676	-0.90	2.88	-0.09	0.01	-0.01	0.08
Total N	2871			1257		1614	

* The distribution of answers to the question on postmaterialism is indicated in the text.

Table 1. Descriptive statistics for the dependent variables

With respect to postmaterialism, very similar distributions were obtained: 36% chose "maintaining order" in the EVS sample, as compared to 37% in the WVS one; "Giving people more say" reaches 21% in the EVS, and 20% in the WVS; 30% in the EVS, and 32% in the WVS chose "Fighting rising prices"; 13% (EVS), respectively 12% (WVS) preferred "protecting freedom of speech" as first choice.

Up to this point, WVS and EVS appear similar. Only the differences in life satisfaction seem troublesome, but they may be due to eventual differences in structure between the two samples, which were not addressed during weighting. I am not aware of such differences, but they might exist. For instance, one sample might have a higher share of employed people as compared to the other one. To avoid such eventuality, each of the four variables was predicted in regression models that controlled for age, gender, education, and employment status, along with an indicator of time, and a dummy for the survey. Multilevel models were set up, considering that respondents are nested in counties (there are 42 counties in Romania). For postmaterialism, multinomial regression was

considered under multilevel assumptions. Stata 15 was employed, using the procedures "mixed" (for continuous outcomes), respectively "gsem" (for multilevel multinomial regression).

To model the time variable, I employ several strategies. I introduce them along the way, presenting the method and then the results for each step of the analysis.

First, time is part of the equation as the mere date of the data collection. This is just a simple way to observe potential trends. In Table 2 the results of this approach are on the row labelled as "Linear" model. Despite initial observation that levels of life satisfaction may increase in the EVS sample, after controlling for various confounders, it turns out that there is no difference between surveys, and also there is no linear effect from time.

The second approach was to replace time with dummies for the month of the year when data were collected. Survey continues to be unimportant, except for the case of religious belief, which is lower in the case of WVS as compared to EVS.

How to model time	DV / effect	**Religious belief**	**Life Satisfaction**	**Materialist/Postmaterialist choices**			**Political Trust**
				More say	Fighting prices	Freedom of speech	
Linear	Time	ns	ns	ns	ns	ns	ns
	Survey	ns	ns	ns	ns	ns	ns
Months	Time	Feb: -0.3* Mar: -0.5*	ns	May: -15.5***	May: 2.5***	May: 1.3*	May: 0.3**
	Survey	-0.4*	ns	ns	ns	ns	ns
Distance to Christmas	Time: days to/from Xmas	ns	ns	ns	ns	ns	ns
	after Christmas	ns	ns	ns	ns	ns	ns
	Survey	ns	ns	ns	ns	ns	ns
Distance to Xmas or Easter	Time: days to/from Xmas	ns	ns	ns	ns	ns	ns
	Time: days to/from Easter	ns	ns	ns	ns	ns	ns
	Before Xmas	ns	ns	ns	ns	ns	ns
	After Easter	ns	ns	ns	ns	ns	ns
	Survey	ns	ns	ns	ns	ns	ns

Surveys: WVS effect as compared to EVS. Month of reference: January. Materialism/Postmaterialism has the reference category: "Maintain order." NS means "not significant." *** p<0.001 ** p<0.01 *p<0.05

Table 2. Main results: impact of time and survey on the dependent variables

The results start showing time-dependent variations in the outcomes. The less important are the ones for materialism/postmaterialism. While data collected in May look significantly different

from those collected other times, since only a few questionnaires were fielded in May, one can overlook this variation. For life satisfaction, no impact was noticed, while for Political trust it is again only the month of May (that is, after Easter) that produces a significant association. For all three variables there seems to be no immediate connection to public holidays, either of religious (red marks in Figure 1) or of non-religious nature.

Things are different with respect to religious beliefs, when controlling for month of the year when data were collected. For this dependent variable, there is a consistent decrease in reported religious belief after Christmas, to stop when Easter is approaching, as illustrated in Figure 2.

The next set of models includes the distance to Christmas in number of days. It is computed as day to/from Christmas, and it is completed by a dummy variable for the questionnaire being completed after Christmas. These models bring no significant effect for time or survey.

The last set of models use measures of time as the distance (number of days) between the moment of data collection and Christmas, respectively Easter. No significant effect is visible this time as well.

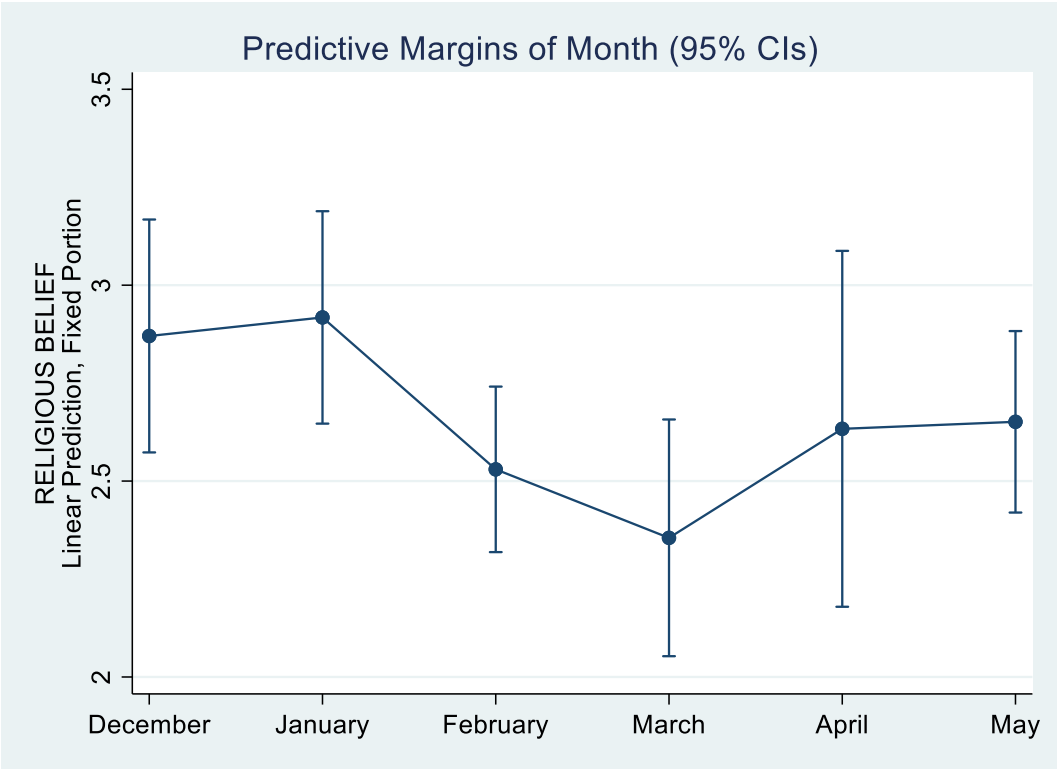


Figure 2. Marginal effects of month on religious belief (based on the second model from Table 2)

Summary

In bivariate analysis, there is no significant difference between the two EVS and WVS surveys for Romania, except for life satisfaction, which is slightly higher in the WVS sample collected over a period including Christmas. The difference in average life satisfaction across the two samples

disappears when controlling for respondents' basic socio-economic characteristics. The lack of significant effects opens the door to legitimately aggregating and comparing surveys, despite the period of the year devoted to data collection.

However, the situation proves to be more complex in multivariate analysis of religious belief. Belief appears to be dependent on the time of the year when data are collected. The analysis shows no significant effect when counting days from Christmas or from Easter. However, in February and March, religious belief decreases quite a lot: 0.3, respectively 0.5 as compared to January. On a scale from 1 to 5, 0.3 means close to 10% of the scale, showing a practical significance that cannot be overlooked. This implies that, for some variables, aggregating data coming from various surveys faces the challenge of similarity with respect to the period of the year when the fieldwork was carried out.

Coming back to the other three dependent variables, the subsequent models depicted in the lower panes of Table 2 provide no significant effect. This was also observed in the first two multivariate models. The immediate conclusion is that in these two studies, time of data collection is irrelevant for life satisfaction, postmaterialism and political trust. It also seems to have no dependency related to the type of scale, since there is no consistent impact for continuous variables.

Time had a significant association with the measurement of religious belief, as shown in the second model. A close look at the variation observed after controlling for various confounders, and the trend becomes clear (Figure 2): Religious belief decreases immediately after winter holidays and increases before and during Eastertime. Surprisingly, despite discussions of seasonality of religious practice (e.g. Olson & Beckworth, 2011), the literature does not include warnings on dynamics of religious belief across the year. This research note provides indications that time of data collection may matter as belief tends to be lower in February and March, that is, between Christmas and Easter, as compared to around Christmas and Easter. Further research should be done in order to inspect the exact radius of high religiousness around the two salient holidays. However, this is beyond the scope of this research-note.

The main goal was to show that survey researchers need to be aware to time-dependency when comparing estimates across different surveys. At least for religiosity-related measures, the period of the fieldwork matters with respect to final estimates. Furthermore, the impacts seem to be independent of the measure that is considered, but further research should inspect other measurement for religious belief. For life satisfaction, an effect of the survey is reported in this study, but it also requires further investigation.

Bogdan Voicu is Research Professor with Romanian Academy (Research Institute for Quality of Life) and Professor of Sociology with 'Lucian Blaga' University of Sibiu. He is principal investigator in Romania for EVS and WVS.

References

Davidov, E. (2010). Testing for comparability of human values across countries and time with the third round of the European Social Survey. *International Journal of Comparative Sociology*, *51*(3), 171-191.

Ehrhardt, J. J., Saris, W. E., & Veenhoven, R. (2000). Stability of life-satisfaction over time. *Journal of Happiness Studies*, *1*(2), 177-205.

Harkness, J. (1999). In pursuit of quality: issues for cross-national survey research. *International Journal of Social Research Methodology*, *2*(2), 125-140.

Inglehart, R., & Abramson, P. R. (1999). Measuring postmaterialism. *American Political Science Review*, *93*(3), 665-677.

Kankaraš, M., Moors, G., & Vermunt, J. K. (2010). Testing for measurement invariance with latent class analysis. *Cross-cultural analysis: Methods and applications*, 359-384.

Olson, P. J., & Beckworth, D. (2011). Religious change and stability: Seasonality in church attendance from the 1940s to the 2000s. *Journal for the Scientific Study of Religion*, *50*(2), 388-396.

Tufiş, C. D. (2012). *Learning Democracy and Market Economy in Post-Communist Romania*. Iaşi: Institutul European.

Voicu, B. (2012). The impact of presidential elections on life satisfaction, pp. 235-258 in Mircea Comşa, Andrei Gheoghiţă, and Claudiu Tufiş, eds. *Romanian presidential elections, 2009,* Cluj: Cluj University Press [in Romanian language]

News

Human Understanding Measured Across National (HUMAN) Surveys: Crowdsourced Data Harmonization

by Andrew James Klassen

HUMAN Surveys is a data harmonization project that formats and merges freely available and nationally representative public opinion surveys using STATA software (A. Klassen, 2018a, 2018b). Following a recent update, the project now includes almost ten million respondents from over 160 countries between 1948 and 2019. Both the data and their codebook are available via Harvard Dataverse. This short article gives a brief overview the HUMAN Surveys project and its database structure, and outlines how HUMAN Surveys, by using crowdsourcing, can represent an alternative, more cost-effective approach to 'traditional' large-scale ex-post survey data harmonization projects.

The original aim of HUMAN Surveys was to investigate how the configuration of political systems, design of governing institutions, and performance of national economies affect public attitudes. It started as a PhD project on perceived electoral integrity and has since been expanded considerably (Klassen, 2014). Harmonized variables currently include satisfaction with democracy, support for democracy, perceived electoral integrity, generalized social trust, and basic demographics.

Figure 1 maps the distribution of the 8,407 country-survey observations currently included, each of which represents one survey conducted in one country.

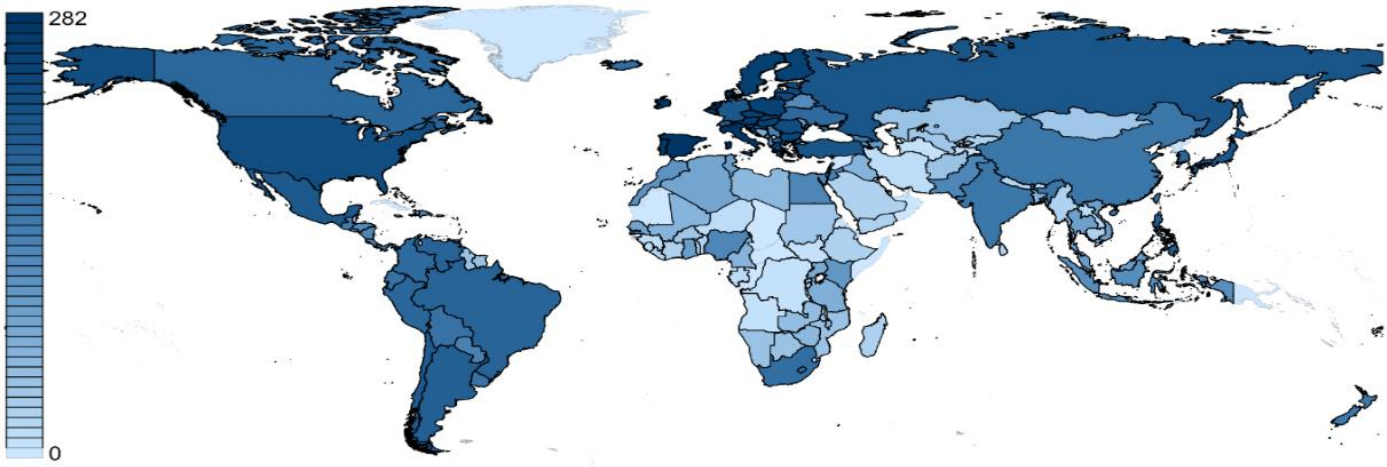


Figure 1. Country-survey heat map of number of surveys conducted per country across all sources.

The HUMAN Survey project uses Stata scripts to format selected individual-level variables, and to produce three databases: of micro-level individual respondents, macro-level country-survey scores, and macro-level country-year scores. Country-survey scores are useful for comparing survey sources, while country-year scores are useful for time-series analyses (for details on the datafiles, see Klassen 2018). Table 1 summarizes respondent data from the 29 survey sources currently included.

In the HUMAN Surveys project, formatted individual-level variables are as close as possible to their original formats. This means that, while a conceptual definition informs the recoding decisions for a given harmonized target variable, as much information from the format of the source variable is retained as possible (for details, see Klassen 2018a). Each target variable is identified by a prefix, followed by a three-digit number. The first number groups variables by survey item, while the two remaining numbers indicate the response scale. Finally, a letter is added at the end of recoded original variables to distinguish between different versions of survey items with the same number of responses (Klassen 2018a, p. 2).

To illustrate, the harmonized measure Social Trust (Common), at_002, distinguishes between respondents that generally trust most people (1=yes) and those who would be more careful (0=be careful). Since original questions on trust differ with respect to wording and scale properties, source item metadata are stored via additional variable variants that researchers can easily identify thanks to multi-digit codes. Table 2 illustrates retained source item variability for the harmonized social trust indicator (column 1), and the recoding decisions (columns 2 & 3). All variables are fully described in the Human Survey project Codebook (Klassen 2018a).

Sources	Years	Respondents
Afrobarometer	1999 - 2015	204,464
American National Election Studies	1948 - 2018	71,489
AmericasBarometer	2004 - 2016	286,313
Arab Barometer	2006 - 2019	69,561
Arab Transformations Project	2013 - 2014	9,809
Asia Barometer	2003 - 2007	46,094
Asian Barometer Survey	2001 - 2016	79,446
Australian Election Study	1987 - 2016	30,597
Australian National Political Attitudes Surveys	1967 - 1979	5,943
Australian Social Cohesion Survey	2007 - 2016	15,780
Australian Survey of Social Attitudes	2003 - 2016	27,396
Caucasus Barometer	2008 - 2017	44,584
Comparative Study of Electoral Systems	1996 - 2018	246,513
Consolidation of Democracy in Central and Eastern Europe	1990 - 2001	27,441
EU Neighbourhood Barometer	2012 - 2014	94,501
Eurobarometer - Applicant and Candidate Countries	2000 - 2004	143,226
Eurobarometer - Central and Eastern	1990 - 1997	125,875
Eurobarometer - Standard and Special	1962 - 2019	4,513,780
European Social Survey	2002 - 2017	372,935
European Values Study	1981 - 2018	188,916
Global Attitudes and Trends	2002 - 2017	485,431
IntUne - Integrated and United	2007 - 2009	34,785
International Social Survey Programme	1985 - 2019	1,213,811
Latinobarómetro	1995 - 2017	410,944
New Europe Barometer	2000 - 2005	76,492
New Russia Barometer	1992 - 2009	34,071
Political Action / Political Ideology	1973 - 1976	12,588
Voice of the People Series	2000 - 2012	481,067
World Values Survey	1981 - 2014	330,354

Table 1. Summary of survey sources, years, and respondents

	Be Careful	Trust Others
at_002a	0	1
at_002b	0	1
at_003a	0 – 1	2
at_004a	0 – 1	2 – 3
at_004b	0 – 1	2 – 3
at_004c	0 – 1	2 – 3
at_004d	0 – 1	2 – 3
at_005a	0 – 2	3 – 4
at_007a	0 – 3	4 – 6
at_010a	0 – 5	6 – 9
at_010b	0 – 5	6 – 9
at_011a	0 – 5	6 – 10

Source: Klassen 2018a, p. 3

Table 2. Recoding of values to create common social trust variable

Current Data Harmonization Projects

Other projects have harmonized public opinion data by paying research assistants, having computer scientists customize software environments, or dedicating hundreds of hours to formatting and managing data. The Survey Data Recycling project, built upon the Democratic Values and Protest Behavior project (e.g. Tomescu-Dubrow & Slomczynski, 2016), uses a suite of customised software tools and paid database specialists to manage the technical aspects of the project (Powałko & Kołczyńska, 2016). The Old and New Boundaries project analyzes different variables from many of the same sources, but similarly requires research assistants to format and manage data (Schnabel, Quandt, & Bechert, 2019). Other harmonization projects have focused on evaluating support for democracy (Claassen, 2019a, 2019b; Foa & Mounk, 2016, 2017), but not expanded beyond this research interest. The huge amount of freely available public opinion data can be expensive and time-consuming to manage. Projects relying on funding to undertake data harmonization are likely to stop doing so when the funding ends, but new rounds and sources of data continue being released every year. Developing complex and customized software environments can also mean that the work of survey harmonization is beyond the capabilities of most social scientists.

Crowdsourced Data Harmonization

The aim of HUMAN Surveys is to provide a platform that lets researchers easily harmonize public opinion data and contribute additional variables and survey sources. The intention is to enable formatting of every variable from every round of freely available survey source. The goal is also to operate indefinitely and without substantial funding while expanding the number of formatted variables and survey sources. This is an ambitious goal because, as Figure 2 shows, the number of national surveys conducted per year increases with time. Note that recent years show a decline

because most survey sources do not release their latest rounds for free immediately.

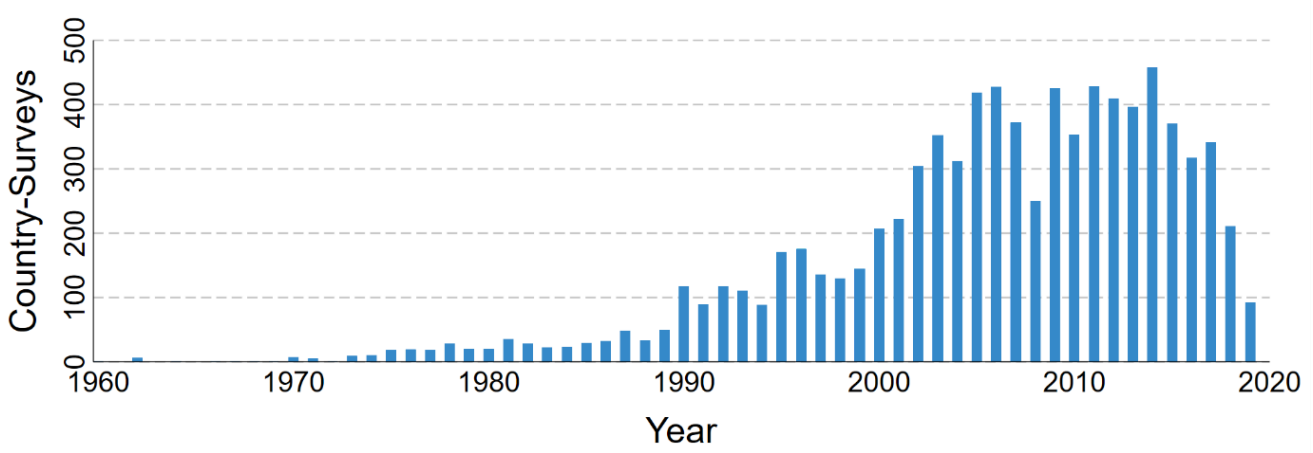


Figure 2. Country-Survey Histogram that shows the total number of national surveys conducted globally per year across all included sources

Comparative public opinion researchers spend countless hours working 'for free' cleaning up survey datasets before analysing them. We separately repeat the same steps on the same datasets, redundantly doing work that others have already done and will repeat in the future. Since many researchers already use Stata to manage their data and conduct analyses, we can crowdsource the cleaning and formatting work already being done by the global public opinion research community.

In the context of this project, 'crowdsourcing' refers to the following process: Multiple researchers and working groups around the world can download the existing scripts and add any new variables and sources that they need for their studies. The main requirement for using the existing scripts would be an agreement to send updated versions with additions back so they can be compiled into the constantly evolving master scripts. Other people and groups would start with updated scripts and add yet more variables and sources that would be subsequently compiled for re-release. HUMAN Surveys would act as an update aggregator and quality control hub, releasing codebooks, standards, guidelines and regularly releasing new scripts with more variables and including more sources (Klassen, 2018). Additions sent back by researchers would then be compiled into updated scripts, which would be re-released as the most recent versions and the process would start all over again. The work of formatting and harmonizing survey data would thereby be outsourced to the global public opinion research community (the crowd).

A crowdsourcing approach takes advantage of the fact that hundreds of researchers around the world have already done similar data harmonization work and thousands more will. However, instead of each person or group starting from scratch, each could start with the latest versions of the

scripts and add any additional variables or sources that they need for their MA thesis, PhD dissertation, research study, or project. Each subsequent project would have more variables and sources to use as a starting point for their research. Crowdsourcing data harmonization makes it possible for researchers with Stata skills to contribute additional variables and sources. The effort can be continued indefinitely with limited funding. The main financial cost would be maintaining a website, such as humansurveys.org, to provide support documentation and a platform to release the updated scripts.

Andrew James Klassen is the Principal Researcher of HUMAN Surveys, a project that uses Stata scripts to harmonize freely available and nationally representative public opinion data.

References

Claassen, C. (2019a). Does Public Support Help Democracy Survive? *American Journal of Political Science*, Early Access. doi: 10.1111/ajps.12452

Claassen, C. (2019b). In the Mood for Democracy? Democratic Support as Thermostatic Opinion. *American Political Science Review*, Early Access, 1-18. doi: 10.1017/S0003055419000558

Foa, R. S., & Mounk, Y. (2016). The Danger of Deconsolidation: The Democratic Disconnect. *Journal of Democracy*, 27(3), 5-17.

Foa, R. S., & Mounk, Y. (2017). The Signs of Deconsolidation. *Journal of Democracy*, 28(1), 5-16.

Klassen, A. (2014). *Perceptions of electoral fairness: public behaviour and institutional design compared across 80 countries.* (PhD), The Australian National University, Canberra, ACT, Australia.

Klassen, A. (2018a). Human Understanding Measured Across National (HUMAN) Surveys: Codebook for Respondent Data (Publication no. doi/10.7910/DVN/QLKR85). from Harvard Dataverse http://dx.doi.org/10.7910/DVN/QLKR85

Klassen, A. (2018b). *Human Understanding Measured Across National (HUMAN) Surveys: Respondent Data* (Publication no. doi/10.7910/DVN/XEA5FD). from Harvard Dataverse http://dx.doi.org/10.7910/DVN/XEA5FD

Klassen, A. J. (2018). "Human Understanding Measured Across National (HUMAN) Surveys: An Introduction." Paper presented at the 2018 APSA Conference, Brisbane, QLD, Australia.

Powałko, P., & Kołczyńska, M. (2016). Working with Data in the Cross-National Survey Harmonization Project: Outline of Programming Decisions. *International Journal of Sociology*, 46(1), 73-80.

Schnabel, A., Quandt, M., & Bechert, I. (2019). *Old and new boundaries: National Identities and Religion.*

Tomescu-Dubrow, I., & Slomczynski, K. M. (2016). Harmonization of Cross-National Survey Projects on Political Behavior: Developing the Analytic Framework of Survey Data Recycling. *International Journal of Sociology*, 46(1), 58-72.

The CLOSER Consortium of UK Longitudinal Studies: Facilitating New Avenues of Research across Studies and over Time

by Dara O'Neill and Rebecca Hardy

The UK has an extensive history of longitudinal research, with the first nationally representative birth cohort commencing in 1946. This has since been followed by many additional cohort and panel studies at the national and regional level. Eight of these studies are part of the CLOSER consortium, alongside the UK Data Service and the British Library. Our interdisciplinary consortium collectively aims to maximise the use, value and impact of longitudinal studies and the research they generate. As part of achieving these aims, CLOSER coordinates a diverse set of harmonisation projects that seek to document the availability and comparability of longitudinal cross-study data, as well as generating new harmonised data resources covering different social and biomedical research domains. This work has been made possible through support from the UK's Economic and Social Research Council (ESRC) and Medical Research Council (MRC).

To date, we have coordinated 16 harmonisation work packages, covering the following topic areas:

For each topic, our work packages have sought to evaluate the similarities and differences in the measurement protocols and instruments used across the CLOSER partner studies as well as within the studies over time. In some instances, such as with the DNA methylation work package, studies from outside the consortium have also been included. To date, half of our work packages have documented these efforts in the form of open access resource reports that are in the process of being disseminated via the main CLOSER website (closer.ac.uk). The remaining eight work packages have identified a sufficient level of equivalence in the data to proceed with the derivation of harmonised variables and these are being made available for wider research use through deposits to the UK Data Service repository (ukdataservice.ac.uk). These deposits are accompanied by detailed user guides as well as code files that are intended, where the source data are also available to researchers, to aid reproducibility. These can help ensure the harmonised variables are maximally adaptable to new research questions, as the code files can enable researchers to fully understand and amend decisions made during the original harmonisation process, as well as to more easily apply similar harmonisation processes to datafiles from other sweeps or from new studies.

To further facilitate the identification and usage of the data available across our partner studies, we are also building a metadata enhancement/search platform that is designed to be a unified point of access to discover information about the questions used and variables collected across the studies (discovery.closer.ac.uk). We also build capacity and best practice in harmonisation through networking and training events and we have developed an online training platform, the CLOSER Learning Hub (learning.closer.ac.uk), to introduce students and early career researchers to the principles and possibilities of longitudinal cross-study research. In addition, we advocate for increased support for new harmonisation efforts through promotion and engagement work with policy makers and funding organisations.

As a consortium, collaborative working is the basis of all our research efforts and we very much welcome additional opportunities to share learning on harmonisation and to engage with more partners in exploring new avenues of longitudinal cross-study research.

For further information about CLOSER, check out our website (closer.ac.uk), see our data resource profile in the *International Journal of Epidemiology* (https://doi.org/10.1093/ije/dyz004) or get in touch with us directly (closer@ucl.ac.uk).

Dara O'Neill is the research theme lead for harmonisation at CLOSER, based at University College London.

Rebecca Hardy is a Professor of Epidemiology and Medical Statistics at University College London and the Director of CLOSER.

Harmonisation Case Study: The Scottish Surveys Core Questions

by Sarah Martin

Scottish Government surveys underpin the evaluation of the National Performance Framework, the co-created purpose and vision for Scotland, and performance dashboard of 11 broad National Outcomes, measured through 81 National Indicators.[1]

Surveys gather information on the demographics, experiences, attitudes, and views of 5.5 million residents, and the results inform and evidence policy decisions. Two hundred interviewers in different contracted agencies are dispatched to nearly 40,000 randomly sampled addresses on even the remotest islands to ensure the data collected are of the highest standard.

Each year, items on lifestyle and health behaviour are asked of 5,000 respondents in the Scottish Health Survey[2], including up to two children and up to the ten adults per household; experiences of crime and victimisation are asked of 6,000 random adults in the Scottish Crime and Justice Survey[3], and travel, housing, energy use, financial, cultural information and more are asked of 10,000 highest income householders or a random adult in the Scottish Household Survey[4]. Together, they result in an annual sample of around 20,000 individual adult interviews, with a core set of common questions, e.g. demographic details.

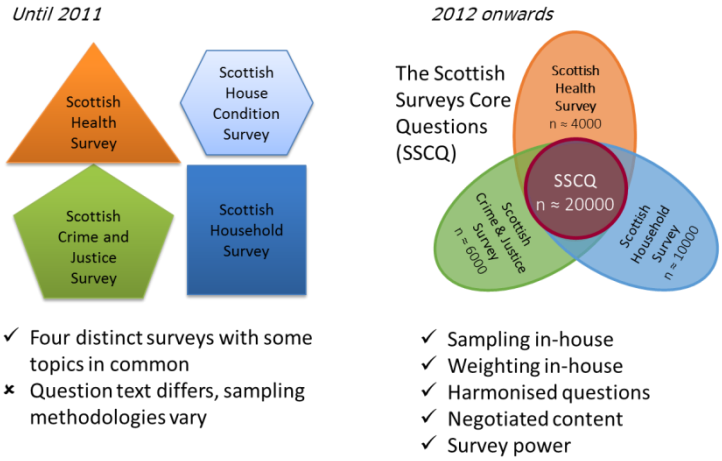
A harmonised and robustly pooled sample that enables analysis of outcomes for small minority groups and local geographies was the dream – and this is its story.

Until the financial downturn of 2008, these surveys were larger, and sat independently as separate commissions in different departments. The Long-term Survey Strategy 2009-2019[5] used the climate of reduced public sector funding to achieve two long-anticipated objectives: harmonisation and coordination across the organisation. Since 2012, this strategy has saved £2.8m annually by

coordinating and rationalising the surveys, but stopped short of creating a single survey with modules as pioneered in Wales three years later.[6] Instead, the Scottish route was harmonisation-based: on the surface somewhat less ambitious, but fundamentally powerful.

In this "best of both worlds" approach, each survey remained as an independent commission and publication, serving its primary purpose to meet the policy needs of its department. Meanwhile, the coordinated questionnaire reviews achieved many compromises and aligned surveys with the 2011 Census demographic questions.[7] Some variables remain a challenge: income, for example, is, for good reasons, still asked and calculated differently, while "caring" estimates are higher in the Household Survey – likely because the question sits alongside a set on volunteering. Estimates of people living with long-term conditions are considerably higher in the Health Survey, which requires a very detailed show card to route to down-stream health questions, while a shorter and simpler form is used in the other two surveys.

Next, each survey contributed key outcome questions, fundamental to Scotland's National Performance Framework, to be asked in all three surveys, forming a flexible set of core questions alongside the harmonised demographic cross-variables. Most recently, all three surveys asked about perception of crime in the local area, general health, mental wellbeing, satisfaction with local services, and more.



The crucial element to enable pooled estimates was control over sampling and weighting, so these were brought in house. With the help of the UK Office for National Statistics methodology unit, rigorous methods were developed, which ensured that addresses are only sampled once across all surveys over four years. A pooled weighting strategy allows us to publish results for the pooled sample as if it originated from a single survey – the Scottish Surveys Core Questions (SSCQ).[8]

Now in its 7th year, SSCQ has produced the first official statistics of sexual orientation minorities[9], analysis of perceptions of police performance by ethnic minority and by latent class analysis[10], up-to-date demographic estimates for non-Census years[11], and outcome estimates down to small areas for local government policy.[12] Crime, health, and mental wellbeing patterns across the population have been analysed in a detail[13] that would have cost upward of £3m per year if commissioned as a separate survey – analysis that is now a free by-product of the harmonisation effort.

Dr Sarah Martin is a Statistician in the Scottish Government and currently Head of Survey Strategy in the Office of the Chief Statistician. Her duties – aside from long-term planning and procurement – include analysis and publication of the Scottish Surveys Core Questions and acting as harmonisation champion for the Government Statistical Service.

---

1 nationalperformance.gov.scot
2 www2.gov.scot/Topics/Statistics/Browse/Health/scottish-health-survey
3 www2.gov.scot/Topics/Statistics/Browse/Crime-Justice/crime-and-justice-survey
4 www2.gov.scot/Topics/Statistics/16002
5 www2.gov.scot/Topics/Statistics/About/SurvStrat
6 gov.wales/national-survey-wales
7 https://www.scotlandscensus.gov.uk/glossary/census-questionnaire-2011
8 www2.gov.scot/Topics/Statistics/About/Surveys/SSCQ
9 www.gov.scot/publications/sexual-orientation-scotland-2017-summary-evidence-base/pages/0/
10 www2.gov.scot/Topics/Statistics/About/Surveys/SSCQ/SSCQ2014/SSCQ2014-PolCon
11 www2.gov.scot/Topics/Statistics/About/Surveys/SSCQ
12 www2.gov.scot/Topics/Statistics/About/Surveys/SSCQ/SSCQ2018
13 www2.gov.scot/Topics/Statistics/About/Surveys/SSCQ/MySSCQ

Reflecting Europeanisation: Cumulative Meta-data of Cross-country Surveys as a Tool for Monitoring European Public Opinion Trends

By Piotr Jabkowski and Piotr Cichocki

Our research team at the Faculty of Sociology (Adam Mickiewicz University, Poznan) is going ahead with a 3-year project funded within the OPUS framework of the National Science Centre, Poland (2018/31/B/HS6/00403). The project will investigate the changing dynamics of attitudes towards European integration through an ex-post analysis of the results of major cross-country surveys since 1999.

Research focus

In Europe, there is an abundance of high-quality cross-country surveys, which translates into an exceptional capacity for monitoring opinion trends over time. Yet, those European projects prove highly diverse in terms of their institutional aims, research focus as well as spatial and temporal coverage. Crucially, they also exhibit marked methodological differences that are bound to impact the quality of gathered data (Kohler, 2007). This heterogeneity hampers substantive cross-project comparisons due to the difficulty of establishing equivalence between measurements. Apart from methodologically focused studies of survey outcomes, e.g., non-response (Groves and Peytcheva, 2008), within-household selection impact (Gaziano, 2005) or data quality assessment (Slomczynski, Powalko and Krauze, 2017), the prevailing norm in empirical studies is for data from a particular project to be analysed in separation from other projects. If such cross-project references feature at

all, it typically happens in the introduction or discussion of results as descriptive reference-points (Heath, Martin and Spreckelsen, 2009).

This project would attempt to produce an integrated cross-project insight into the dynamic of attitudes towards European integration, focusing on the main multi-wave pan-European surveys implemented since the late 1990s. Over the last two decades, cross-country surveys have proliferated with the establishment of the European Social Survey (ESS) and the extension of country coverage of the International Social Survey Programme (ISSP). This period has also been transformative for European integration due to the impact of EU enlargement as well as of a series of unprecedented political, social and economic crises.

The principal aim of this research project is to conduct comparative analyses of selected measurement models that relate to European identification and legitimacy of European institutions over time between the survey results of different cross-country surveys. This translates into two specific aims: (A) comparison and cross-project validation of measurement models based on survey-derived indicators, (B) evaluating the impact of methodological characteristics on the quality of survey results. In order to pursue those aims, it would be necessary to build cumulative databases at three levels of aggregation: (1) micro – variable values and weights, (2) mezzo – question-level methodological characteristics, (3) macro – survey-level methodological characteristics.

The project is guided by three broad research hypotheses.

H1: Cross-project replicability of published results of substantive empirical studies is low due to methodological incompatibilities of indicators used to construct measurement models, correlates, and control variables.

H2: Methodological characteristics of question-items are likely to have a much stronger impact on the quality of measurement models than the macro-characteristics of surveys.

H3: High public-anxiety external events, such as the economic or migration crises, exert a uniform impact across all survey projects on the indicators of European identification and legitimacy of European institutions as well as their correlates (covariates and main factors).

Empirical scope

Our data-query would encompass projects fulfilling the following criteria: (1) having a comparative focus, (2) being non-commercial, (3) having an academic renown demonstrated by the high number of publications in the Web of Science Core Collection (WoS), (4) having a longitudinal time-perspective, (5) having nationally representative samples, (6) allowing for open and free access to databases and research documentation. Given such criteria, four survey-projects have been chosen: Eurobarometer (EB, conducted biannually since 1974), European Values Study (EVS, conducted

once per decade since 1981), ISSP (conducted annually since 1985), and the ESS (conducted biennially since 2002).

Project name	**Time frame**	**Number of waves**	**No. of surveys**
Eurobarometer	2001-2018	19	≈1000
European Social Survey	2002-2018	9	≈220
European Values Study	1999-2017	3	≈120
International Social Survey Programme [a]	2003-2014	4	≈100

[a] National Identity II (2003), Citizenship (2004), National Identity III (2013), Citizenship (2014).

Table 1. Survey projects selected for comparative consideration

While the four selected survey-projects share common characteristics, they also differ substantially when it comes to the questions-items they contain and how these items are asked. Thus, apart from the socio-demographic variables, their data-structures would not allow for straightforward harmonization. Given that they all touch upon topics of European identification and the legitimacy of European institutions, this common denominator would allow ample space for cross-project comparisons of measurement models.

Methodological approach

Harmonisation projects inspire our approach, but this project is different from them in several respects. First and foremost, instead of aiming at reducing original variables towards one common denominator, we will produce a cumulative database retaining specific qualities of measurement of particular projects. Apart from socio-demographic characteristics and weights, which are going to be harmonised, all the substantive variables pertaining to European identification and legitimacy of European institutions will retain their original form. Comparative analysis, which might be referred to as 2nd-degree harmonisation, will be performed concerning measurement models (indices, scales) identified in published empirical studies. Although comprehensive harmonisation is not our aim, harmonisation procedures will apply to variables that (1) describe socio-demographic characteristics of respondents and (2) provide information necessary for conducting external and internal evaluations of survey quality.

Furthermore, the project would also yield meta-bases containing a set of methodological control variables describing the characteristics of indicators (mezzo) and country-surveys (macro). This would enable an investigation of whether the peculiarities of measurement influence the analyses of survey results (e.g., length of scales, ways of coding missing data) as well as of fieldwork execution (e.g., sample frame, fieldwork procedures). All databases produced as well as

methodological documentation accumulated throughout this project are going to be put in the public domain available for re-use for free by registered users.

The three respective databases would have the following characteristics: (A) The micro-base will accumulate all values of selected indicators for individual observations (approx. 1.5 million respondents) derived from all the relevant research waves of the four projects (approx. 1.4 thousand surveys). The database will contain two components: (i) the cumulative, i.e., original values of selected indicators pertaining to attitudes towards European integration, (ii) the harmonized, i.e., transformed values of socio-demographic variables and weights; (B) The mezzo-base will provide an inventory of methodological characteristics for all selected basic indicators; and (C) The macro-base will provide measurement characteristics (including internal and external measures of sample quality) for all surveys under consideration.

The information stored in the mezzo-base will allow for validation of substantive indicators by reference to control variables. In comparison to the common practices within current harmonization projects, this approach is innovative as they typically only include macro-level characteristics of surveys. The inclusion of methodological characteristics of original questions in the analysis constitutes the necessary effect of the decision to eschew direct harmonization of basic substantive indicators. The set of control variables at the mezzo-level would include: (1) number of question-items underlying the output variable, (2) number of answer-options that the respondent was presented with, (3) ordering of answer-options, (4) coding of item non-response, and (5) item non-response rate.

At the macro-level, control variables would include the typical range of characteristics included in harmonization projects, which have been presented in Table 1. On top of that, the macro-characteristics would also include measures of survey quality. Two representativeness criteria are to be implemented, which are derived from the paradigm of Total Survey Error (Biemer, 2010): (A) external criteria, which refer survey outcomes to known population characteristics (Groves, 2006), (B) internal criteria, which are based on the comparison of distributions of gender in the subsample of heterosexual two-person households against the aprioristic parameter of 50/50 (Sodeur, 1997).

Acknowledgement: This research was funded by the National Science Centre, Poland (2018/31/B/HS6/00403).

Piotr Jabkowski is a professor at the Faculty of Sociology of Adam Mickiewicz University in Poznan and a member of Sampling and Weighting Experts Panel in the European Social Survey.

Piotr Cichocki is an adjunct researcher at the faculty of Sociology of Adam Mickiewicz University in Poznan.

References

Biemer, P. P. (2010) Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5): 817-848.

Gaziano, C. (2005) Comparative analysis of within-household respondent selection techniques. *Public Opinion Quarterly*, 69(1): 124-157.

Groves, R. M. (2006) Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, 70(5): 646-675.

Groves, R. M. and E. Peytcheva (2008) The impact of nonresponse rates on nonresponse bias a meta-analysis. *Public Opinion Quarterly*, 72(2): 167-189.

Heath, A., J. Martin and T. Spreckelsen (2009) Cross-national Comparability of Survey Attitude Measures. *International Journal of Public Opinion Research*, 21(3): 293-315.

Kohler, U. (2007) Surveys from Inside: an Assessment of Unit Nonresponse Bias with Internal Criteria. *Survey Research Methods*, 1(2): 55-67.

Slomczynski, K. M., P. Powalko and T. Krauze (2017) Non-unique records in international survey projects: the need for extending data quality control. *Survey Research Methods*, 11(1): 1-16.

Sodeur, W. (1997) Interne kriterien zur Beurteilung von Wahrscheinlichkeitsauswahlen. *ZA-Information/Zentralarchiv für Empirische Sozialforschung*, (41): 58-82.

The Cross-National Biographies - Young (CNB-Young) Project: Harmonizing Panel Data for the Study of Youth Employment Precarity

By Anna Kiersztyn

The "Dynamics of youth employment precarity: drivers, trajectories, and outcomes in a cross-national perspective" project funded the National Science Centre, Poland research grant (2018/31/B/HS6/02043), led by Anna Kiersztyn and Zbigniew Sawiński, studies the employment trajectories of young adults, with a focus on labour market precarity and its effect on the social structure in a cross-national comparative perspective[1]. This project will create a new panel dataset Cross National Biographies – Young (CNB-Young), which will include harmonized career data on individuals whp are up to 35 years of age from Poland, Germany, Great Britain, and the USA. It will

[1] The project has started on October 1, 2019.

be the first cross-national quantitative dataset covering full employment histories of respondents starting from their first job, their education, changes in household composition, income, and health / well-being.

CNB-Young will harmonize data from four panel surveys: The Polish Panel Survey (POLPAN), the German Socio-economic Panel (SOEP), the U.K. Household Longitudinal Survey – Understanding Society (UKHLS), and the U.S. National Longitudinal Survey of Youth (NLSY79) Young Adults Study. To enable a systematic analysis of cross-country differences, the project will gather relevant contextual data characterising the legal regulations and institutional settings of the four countries under study. The project's main outcomes, including the CNB-Young dataset and documentation, will be made available to the international academic community.

The project is based at the Institute of Sociology, University of Warsaw and the Institute of Philosophy and Sociology, Polish Academy of Sciences. The research team will collaborate with all the survey projects included in CNB-Young. With regard to the methodology of ex-post harmonization of cross-national survey data, CNB-Young builds on the experience generated in the Survey Data Recycling (SDR) project, while also extending the methodology developed in SDR for application to the harmonization of panel surveys.

Substantively, CNB-Young will contribute to the international scholarly debate on the precarization of employment, conceptualized as a career pattern involving short spells of recurrent non-standard employment separated by periods of joblessness, coupled with low and / or unstable incomes. The study will produce new and policy-relevant knowledge on how the interplay of various individual characteristics and institutional factors affect workers' chances of moving into secure employment, or mitigate the possible negative life-course outcomes of early career instability. The harmonized panel dataset will, for the first time, enable a comparative analysis of such patterns and relationships from the perspective of multi-year trajectories.

Methodologically, CNB-Young moves beyond the existing large-scale ex-post harmonization efforts, which either concern cross-sectional data or contain only a limited set of harmonized panel variables with respect to employment or educational history. In the process of its implementation, this project will generate new knowledge on how to address the many methodological challenges involved in the ex-post harmonization of biographical data from single-country surveys.

For additional information on the CNB-Young project, please contact Anna Kiersztyn at chaber@is.uw.edu.pl.

Acknowledgement: This research was funded by the National Science Centre, Poland (2018/31/B/HS6/02043).

Anna Kiersztyn is an associate professor at the Institute of Sociology, University of Warsaw, and a long-standing member of the POLPAN survey team. Her research has used longitudinal data to study the dynamics of underemployment and employment precarity, in particular among young workers.

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. Send it to the co-editors, Irina Tomescu-Dubrow tomescu.1@osu.edu and Joshua K. Dubrow, 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.

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 was a cross-national survey data harmonization project financed by the Polish National Science Centre in the framework of the Harmonia grant competition (2012/06/M/HS6/00322). This newsletter is now funded, in part, by the US National Science Foundation (NSF) under the project, "Survey Data Recycling: New Analytic Framework, Integrated Database, and Tools for Cross-national Social, Behavioral and Economic Research" (SDR project - PTE Federal award 1738502). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation. The SDR project is a joint project of OSU and PAN. For more information, please visit asc.ohio-state.edu/dataharmonization.

Copyright Information