Building Multi-Source Databases for Comparative Analyses

December 16 - 20, 2019

Institute of Philosophy and Sociology, Polish Academy of Sciences

TITLES AND ABSTRACTS

Monday, December 16

The Effects of Data Harmonization on the Survey Research Process

Peter Granda, University of Michigan

From the earliest days of academic-based survey research, the concept of data harmonization and the development of practical implementation tools were central elements in the production of valuable data resources for secondary analyses. It is only now, after more than fifty years, that the importance of data harmonization is clearly recognized.

This presentation will describe some key moments in the early history of data harmonization attempts and their importance for research in both the social and medical sciences. It will highlight more recent projects and innovations that produced major new sources of data. But data harmonization concepts and strategies have affected many parts of the entire research process and the presentation will focus much of its attention on such topics and questions as:

- What is the role of data harmonization in the survey lifecycle?
- How have the different general types of harmonization (input and output) influenced the planning of new surveys?
- Most importantly, where is data harmonization work moving in the future?
- Will the increasing use of social media and other "organic" datasets encourage new data harmonization techniques?
- Can we speculate about the role that AI and other automation tools may have in this space?

Harmonizing Census Data: IPUMS International

Matthew Sobek, Institute for Social Research and Data Innovation, University of Minnesota

IPUMS International harmonizes and disseminates census microdata collected over multiple decades by roughly 100 countries. There is little commonality in the source material over time within countries and no consistency at all across countries. To manage this heterogeneity, IPUMS has developed an extensive data infrastructure driven by metadata. Researchers manage correspondence tables to assign disparate input codes into a common global classification scheme for each categorical variable. These self-documenting tables govern the data harmonization software and provide the value labels for the web dissemination system and generation of statistical package syntax. IPUMS does not transform data using code, unless complex logic is required. Stemming from a single metadata source, the data always stay in sync with the web dissemination system. Other metadata components describe the input data and govern the display of samples and variables in the web system. The development of infrastructure driven by metadata empowers the research staff who best understand the data to accomplish the vast majority of the tasks required for harmonization. Because of the complexity and scale of data harmonized across so many sources, a sophisticated dissemination system is essential, and is an integral part of our approach. Harmonized data are more complicated than data from a single source, and it is essential to convey information without overwhelming researchers.

Harmonizing Time-Use Data: Current Practices and Challenges in the Field

Ewa Jarosz, Institute of Philosophy and Sociology, Polish Academy of Sciences, Sarah Flood, IPUMS, and Margarita Vega Rapun, University College London

Harmonization of time-use surveys requires consistent coding of questionnaire items like in other surveys, but also entails imposing consistency across time diary matrices, which add another dimension of complexity to harmonization. Sequential time diary data can be structured in multiple ways. The time diary design may also vary, and these variations affect the key variables. In Europe time-use surveys are most commonly coded in a format of activity slots of equal duration. In the United States, duration of activity episode is not fixed and may be as short as one minute. The

variability of coding schemes is even greater if non-Western countries are involved. Harmonization requires standardization of such differences. As the process of harmonization also renders culturally-specific activities in the format of a harmonized list of codes, some of the nuance present in the original data is inevitably lost. Importantly, the present standards of harmonization are better fit to Western societies. The purpose of harmonization of time-use data is to compare across time and place. Harmonizing time-use data requires a balancing act between preserving most of the information included in the very rich original dataset and providing the end user with the possibly most universal dataset they may use to fit their purpose. Time-use data may be used to analyze a plethora of social, economic, environmental or biological phenomena from gender inequalities in parental time with children to variability in circadian rhythms and the effect of weather on household energy use. The broad applicability of harmonized time-use datasets makes its harmonization particularly challenging.

Statistical Issues in Analyzing Harmonized Data

Claire Durand, University of Montreal

This presentation aims at synthetizing the main challenges faced by researchers who analyze combined survey data sets. Four different issues and challenges will be addressed.

First, there is a "missing value" issue both at the survey and at the individual levels. Not all surveys ask the same questions pertaining to specific topics. The current answer to this problem is to select only the questions that are asked in the same way in a large proportion of the surveys. This restricts the scope of our analyses. We suggest that repeated measures analysis and item response theory can help use all the available data. With combined data sets, we can also use the information provided by the longitudinal and multi-level nature of the data to improve imputation at the individual level.

Second, in many cases, data are collected repeatedly over time. Researchers tend to describe change over time using bivariate graphs of yearly means or forget about the longitudinal nature of the data altogether. We suggest that it is essential to take the longitudinal nature of the data into account using local regression, trajectory analysis and introducing time in our regression models.

Third, using data from different survey projects constitutes an advantage in terms of statistical power. However, survey projects may not always be conducted in the same country-year

so that it is difficult to assess whether differences are due to methods or to true change. We need to control for these methodological differences in order to make sure that our conclusions are not due to methodological bias.

Fourth, we use data that include weights at the individual level. We may also think about the weights that could be applied at the country level. Should we or shouldn't apply weights in our analyses. Does it make a difference?

Tuesday, December 17

Effecting Rigorous Data Harmonization and Documentation to Understand Data Heterogeneity and Quality

Isabel Fortier and Tina Wey, the Maelstrom Project, Research Institute of McGill University Health Centre

Background

Harmonization of data from different studies is of increasing importance, but it also presents major challenges. A rigorous process of data harmonization requires careful consideration and documentation of factors that can affect heterogeneity and data quality throughout the process. We present challenges, considerations, and approaches for understanding data heterogeneity and quality throughout the harmonization process, based on experience at Maelstrom Research.

Methods

Our harmonization process and documentation follow Maelstrom Research's guidelines for rigorous data harmonization. We examine how factors such as study design, data collection and management procedures, and harmonization processing can affect the results of harmonization. These considerations are illustrated with two case examples of large data harmonization initiatives in ageing cohorts: one prospective and one retrospective.

Results

Factors that should be documented and considered in assessing harmonized results include cohort-specific designs and protocols, study-specific data collection and management procedures, and

harmonization process (definition of target variables, evaluation of harmonization potential, data processing decisions). Important aspects of harmonized data, such as harmonization potential, participant distributions, and missing values, can be affected by both study-specific factors and harmonization processing. Prospectively planned harmonization work generally resulted in higher quality output, in terms of greater harmonization potential across datasets and fewer missing values.

Conclusions

Understanding and using results of data harmonization require detailed understanding of source data and the harmonization process. Careful consideration and transparent documentation of study-specific data and harmonization process, along with iterative evaluation and validation at all stages, is needed to enable data interpretability and inform analyses.

Harmonization of Panel Data: CNEF

Dean Lillard, Department of Human Sciences, Ohio State University, Deutsches Institut für Wirtschaftsforschung, Berlin, and National Bureau of Economic Research, Cambridge, MA

This article describes the past, current, and anticipated future of the Cross-National Equivalent File project (CNEF) – a project that harmonizes data from ongoing longitudinal surveys in nine countries.[1] CNEF is a compendium of data from general population household-based panels in Australia, Canada, Germany, Japan, Russia, South Korea, Switzerland, United Kingdom, and United States. I introduce CNEF – its genesis, the research infrastructure it operates on, and the structure of the CNEF database. Next, I discuss major issues specific to the project's cross-national harmonization of panel data. Regarding substantive variables, we applied harmonization procedures to concepts that have clear theoretical definitions, are measured in objective units and are, in principle, independent of country-specific culture. Hence, CNEF harmonizes data to empirically measure two types of conceptual variables: (a) concepts that are defined by conditions in the objective physical world (e.g. age, biological sex, pregnancy, etc.), and (b) abstract concepts that can be defined rather precisely and are measured in objective units, such as income (household and personal), occupation, and health, among others. While harmonization involves similar issues in cross-sectional and longitudinal data, an obvious difference arises when defining the technical

variables that identify individuals and the households in which they reside over time. I briefly review the rationale for the rule CNEF chose to use. In discussing CNEF's harmonization method, I reflect on (i) lessons learned, for example the extent to which initial considerations changed as CNEF evolved, implications of ex-post harmonization decisions for ex-ante harmonization in subsequent waves of CNEF panel members, and (ii) challenges pertaining both to harmonization *per se*, and its documentation. I conclude with a brief outline of recommendations for researchers interested in harmonizing cross-national panel survey data ex-post.

[1]The Canadian Survey of Labor and Income Dynamics covers the period 1992-2009. Although the British Household Panel Study officially ended in 2008, the UK Longitudinal Household Study (Understanding Society) includes most of the extant BHPS sample.

Theory Driven Approach to Data Harmonization: The True European Voter Project

Herman Schmitt, University of Manheim, Paolo Segatti, University of Milan, and Cees van der Eijk, University of Nottingham

True European Voter (TEV) was a five year Cost Action IS0806 whose aim was to analyse how contextual properties moderate the effects of individual (micro-level) determinants of electoral participation and party choice. The project has been a cooperative effort of five dozens of electoral researchers from almost all European Countries, including well established as well new democracies. The project assembled individual data of National Election Studies conducted in 22 countries and 155 elections from the late 50 to the early years of the current decade. This effort has required a strategy of harmonisation and integration of data which originally were collected by different communities of electoral researchers and according to their different national traditions and scientific interests. Moreover, since the main goal was to study the conditional contextual effects not only on the decision to vote but also on the choice who to vote, TEV project has required a restructuring of the pooled data matrix transforming into in the so-called stacked form. While assessing the effects of the context variations on decision to vote across countries and elections is straightforward, analysis similar impact of vote choice need to solve the major obstacle in comparative chapter, the discrete nature of the voting options in offer, which might vary in number and as to the attributes they have across countries and in some cases also across elections. The presentation illustrates the procedure we followed in the harmonisation process and the rationale of our post harmonisation decisions. We show an example of the scientific potential of TEV dataset

on the basis of a chapter of the book -*The consequences of the context*- we are finalising thanks to the cooperation of more than 20 colleagues.

Analyzing Repeated-Wave Panel Data to Identify Causal Direction: Lessons from a Meta-Analysis of Digital Media Use and Political Participation

Jennifer Oser, Ben-Gurion University of the Negev, Israel

Analysis of repeated-wave panel data is an important methodological approach for determining temporal order in the causal direction of the relationship between key variables. As data harmonization of large cross-national data sets breaks new ground to allow for analysis of greater breadth of data, an important question that emerges from recent social science trends is whether this harmonized data can be leveraged to gain insight on causal inference. I first address this question by briefly review the findings of a meta-analysis study that investigates the temporal causal order in the relationship between digital media use and political participation (Oser, Jennifer & Boulianne, Shelley, forthcoming in Public Opinion Quarterly). The findings of the POQ study, based on 38 survey-based, repeated-wave panel studies (279 coefficients) contradict common assumptions in the literature by showing a reinforcement effect, whereby those who are already politically active are motivated to use digital media. I then proceed to discuss how our research on the POQ article laid the groundwork for a grant application on "Political Efficacy in the Digital Era" that includes a work plan for harmonizing political efficacy variables in cross-national datasets. I conclude with two main observations of potential synergy between meta-analytic and data harmonization techniques, namely (1) The identification of theoretically important social science indicators for which harmonized data are not yet available, and (2) The identification of repeated-wave longitudinal data that can be analyzed in concert with harmonized cross-sectional data to enhance researchers' ability to assess underlying mechanisms in the relationship between key variables.

HUMAN Surveys: Challenges of Formatting Multiple Sources Using Stata Scripts

Andrew Klassen, the HUMAN Surveys Project

Human Understanding Measured Across National (HUMAN) Surveys is a harmonization project that formats and merges freely available and nationally representative public opinion data using Stata scripts. The scripts create metadata, format selected variables, save intermediate datasets, manage data workflows, construct common variables, generate macro-level indices, and produce three data warehouses for analysis and reporting at different levels. The aim of HUMAN Surveys is to enable researchers from around the world to format any variable from any round of any publicly available survey source, and to do so indefinitely without substantial funding. Using Stata as the harmonization tool is advantageous because many public opinion researchers already use this software, providing a familiar platform for these individuals to easily contribute additional variables and survey sources. Despite the familiarity and convenience, there are still many challenges of using Stata to harmonize data. The scripts currently comprise 6000 pages and can be difficult to navigate, contributing additional variables is time-consuming, minor mistakes can stop them executing or substantially alter results, and running all the scripts takes days. This presentation will outline procedures developed to help maintain data quality, facilitate script navigation, and create more efficient workflows. It will also discuss how HUMAN Surveys is building training materials and developing systems that will enable others to use the existing scripts and contribute to their expansion by harmonizing more variables or incorporating more survey sources.

Using Latent Class Analysis for Testing the Equivalence of Multi-Item Scales across Data from Different Survey Programs. The Example of National Identity Types

Markus Quandt, GESIS, and Antonia May, GESIS

The subjective framing of "National Identity" is often distinguished into "civic" and "ethnic" types. Ever since survey-based measures of National Identity types were first analyzed, it is well established that most populations are characterized by a mix of types. This invites the classification of whole nations by the proportion in which particular types are prevalent. However, the coverage

of countries at a given time is often incomplete if using only a single comparative survey. Coverage can in principle be improved if we draw on "similar but not identical" measures from different survey programs, which together cover a larger set of national samples. The presentation will evaluate to what extent this approach is viable, comparing measures from three different comparative surveys - European Values Study 2008, ISSP 2013, and InTune 2007/2009 - for those countries which they jointly cover. As we avail of a multi-item measure - in contrast to most cross-survey harmonization work, which only uses single-item measures, - we have the opportunity to assess measurement equivalence by more formal methods, in this case by equivalence tests that apply to Latent Class Analyses. Preliminary results indicate that we can expect partly valid comparisons when looking at country distributions of National Identity types.

Comparative Sample Quality: Facing the Obstacles to Applying External and Internal Evaluation Criteria to Large Harmonized Datasets

Piotr Jabkowski and Piotr Cichocki, University of Poznan

Large-scale ex-post survey data harmonization holds great promise for the study of sample quality, as it leads to direct comparability of survey outcomes across different projects, waves and countries. However, harmonized datasets involve specific methodological challenges which do not arise in the context of single-project evaluations of sampling quality. For instance, of the five approaches proposed by Groves (2006) only the gold-standard evaluation is typically feasible for harmonized data, while RR subgroup comparisons and analyzing post-survey adjustments work only sometimes. Our presentation will focus on investigating such problems harmonization-specific challenges in evaluations based on the distribution of gender. We will demonstrate that the main stumbling block for external evaluations comes not in finding reliable sources of gold-standards but in the uncertainty about the weights applied in individual surveys. While most cross-country surveys include weights of some kind, design and post-stratification weights are only rarely provided, and in all too many surveys, the available documentation does not allow for a precise determination of what a ‘total weight’ means. We will also demonstrate that an alternative approach to evaluation employing internal criteria (Kohler, 2007; Sodeur, 1997) constitutes some way out of weight-uncertainty.

Friday, December 20

Aggregating Survey Data on the National Level for Indexing Trust in Public Institutions

Joonghyun Kwak, The Ohio State University, and Kazimierz M. Slomczynski, IFiS PAN & CONSIRT

Summary statistics derived from national surveys are commonly used as macro indicators in cross-national comparative analyses, but generally with little or no attention to inter-survey variability stemming from, for example, survey item formulations, or data quality, broadly understood. This study presents a novel approach that the SDR analytic framework motivated, whereby aggregation of individual-level data into country-level indicators accounts for given methodological errors and biases that national surveys display. We use three kinds of information from the SDR database version 1.1. These include (a) the harmonized individual-level measures of trust in three public institutions—parliament, the legal system, and political parties— available in 211 national surveys fielded between 2008-2013 in 57 countries, as part of 18 international survey projects; (b) the harmonization control variables for the trust measures; and (c) a set of control variables that capture survey quality as reflected in documentation, data processing, and computer data files. To obtain best estimates of mean trust values for country-years, we apply linear regression models that control for the lagged effects of trust, methodological differences in survey-specific item formulations (i.e. harmonization controls), and variation in survey quality. Our approach to survey data aggregation contributes to survey methodology for indexing country-level measures of political attitudes and behaviors using individual-level data from cross-national surveys.

Measuring Inequality in Party Representation across Nations and Time using Survey and Administrative Data

Joshua K. Dubrow, IFiS PAN & CONSIRT, and Olga Zelinska, Graduate School for Social Research, IFiS PAN

Party representation is a foremost form of voice in modern democracy as parties and parliamentarians are charged with the responsibility to carry the voice of "the people" into the

legislature. Yet, social groups may feel that their interests are not well represented by the major political parties, while someone else's voice is better heard within the parliament. Social scientists have devised various ways to measure representation across nations and time. Vote-seat share from administrative data and issue congruence, featuring left-right scales from surveys and information from party manifestos, are two of the most often used measures. In addition, representation is often featured as a dimension of democracy in various cross-national democracy-measuring projects, such as in Varieties of Democracy and Democracy Barometer. Our intention is to introduce and develop a new and different measure of party representation that combines survey and administrative data and focuses on the perception of party representation by social groups (gender, age, and education). The survey data come from the European Social Survey (ESS) 2002 - 2016 that contain items on the party that the respondents said they voted for in the last parliamentary election. We combine the ESS item with ParlGov's data on the percentage of parliamentarians in each party in parliament to create a Dissimilarity Index as a measure of inequality in party representation. We present methodological issues that arose in the collection of the data and calculation of the index.

Measuring Protest Trends: A Method for Assessing Macro Tends in Protest Participation

J. Craig Jenkins and Joonghyun Kwak, The Ohio State University

Multiple arguments have been advanced about medium-term macro trends in protest participation (the social movement society thesis, protest apathy in new democracies, restricted opportunities in autocracies) but we lack high quality cross-national data for two or more decades needed to assess these ideas. The best coverage stems from the "have done *ever*" or "lifetime" measure of participation in demonstrations and petitions derived from national surveys but this lacks information about when such participation occurred. We advance a measure based on the responses to this survey question of those aged 18-21 in international surveys harmonized by the Survey Data Recycling Project, which provides us with 36 countries from all major world regions with coverage of 20 years or more, and validate these by comparison with survey responses about participation "in the last 12 months" and the PEA39sixteen project event data measures for 30 European countries. Using this "youth" measure to plot multi-decade trends in participation in

demonstrations, we find mixed evidence for the "social movement society" thesis, general support for protest apathy in the new democracies and support for lower protest rates in autocracies. Country-specific political dynamics seem important alongside these protest trend ideas.

Complexity in Society: From Indicators Construction to their Synthesis

Filomena Maggino, Sapienza Università di Roma

Acknowledgements

This event is organized by CONSIRT (consirt.osu.edu) of The Ohio State University and the Polish Academy of Sciences, with funding from the National Science Foundation for the project Survey Data Recycling (SDR) (PTE Federal award 1738502) and the National Science Centre, Poland, for the project Political Voice and Economic Inequality across Nations and Time (POLINQ) (NCN 2016/23/B/HS6/03916). The event benefits from organizational support that the Institute of Philosophy and Sociology, Polish Academy of Sciences (ifispan.pl/en) provides.